Technical Privacy Metrics: a Systematic Survey

Isabel Wagner
iw@ieee.org
De Montfort University, Leicester, UK

David Eckhoff
de@cs.fau.de
University of Erlangen, Germany

Abstract

The goal of privacy metrics is to measure the degree of privacy enjoyed by users in a system and the amount of protection offered by privacy-enhancing technologies. In this way, privacy metrics contribute to improving user privacy in the digital world. The diversity and complexity of privacy metrics in the literature makes an informed choice of metrics challenging. As a result, redundant new metrics are proposed frequently, and privacy studies are often incomparable. In this survey we alleviate these problems by structuring the landscape of privacy metrics. For this we explain and discuss a selection of over eighty privacy metrics and introduce a categorization based on the aspect of privacy they measure, their required inputs, and the type of data that needs protection. In addition, we present a method on how to choose privacy metrics based on eight questions that help identify the right privacy metrics for a given scenario, and highlight topics where additional work on privacy metrics is needed. Our survey spans multiple privacy domains and can be understood as a general framework for privacy measurement.

1 Introduction

Privacy is a fundamental human right codified in the European Convention on Human Rights, which states that everyone "has the right to respect for his private and family life, his home and his correspondence" [Council of Europe 2010, Art. 8]. However, it is difficult to define what exactly privacy is. As early as 1967, Westin [1967] defined privacy as "the ability of an individual to control the terms under which personal information is acquired and used." Personal information, according to the EU privacy directive, is "any information relating to an [...] identifiable person" [European Parliament 1995]. Nissenbaum [2004] makes these definitions more practical and defines privacy in terms of contextual integrity, where information is associated with a specific context (e.g., a hospital visit), and social norms for this context dictate how information may be used or shared. A privacy violation is then the use of personal information other than the norm allows. Although contextual integrity clearly defines when a privacy violation has occurred, it provides no protection mechanism other than policy and regulations.

Privacy enhancing technologies (PETs) protect privacy based on technology rather than policy, and can thus offer much stronger protection. To judge the efficacy of PETs, privacy metrics are needed that can measure the level of privacy provided by a given PET. Despite the large number of metrics in the literature, a structured and comprehensive overview of privacy metrics does not yet exist. This makes informed decisions about which metrics to select for the evaluation of PETs difficult. In this paper, we structure the landscape of privacy metrics, focusing on technical metrics that measure the effectiveness of PETs. In detail, our contributions are as follows:

- We review conditions for the quality of privacy metrics (Section 2). These are essential as a basis for an informed decision about privacy metrics.
- We describe a selection of privacy domains including communication systems and databases to provide context and examples throughout the survey (Section 3).
- We identify four common characteristics that can classify privacy metrics (Section 4):
  - *Adversary models* describe the capabilities the adversary is assumed to have.
  - *Data sources* describe how the adversary might obtain the information a PET aims to protect: from public data, observable data, re-purposed data, or other sources.

  - *Inputs* describe what information is used to compute a metric: the adversary's estimate, resources available to the adversary, the ground truth, prior knowledge, and parameters.
  - *Output Measures* describe the properties that are measured by privacy metrics to gain an estimate of privacy. Our taxonomy introduces eight categories: a) uncertainty, b) information gain or loss, c) similarity/diversity, d) indistinguishability, e) adversary's success probability, f) error, g) time, and h) accuracy/precision.

- We describe and classify over eighty privacy metrics in Section 5. We focus our selection on popular metrics (in terms of citations) and metrics we found conceptually promising. Where possible, we unify and simplify metric notation and, when appropriate, we discuss advantages and disadvantages of metrics as well as application scenarios.

- We give recommendations on how to choose privacy metrics in Section 6. We structure our recommendations along a series of questions, answers to which will highlight particularly suitable metrics and narrow down the number of candidates.

- We identify areas for future work in Section 7. In particular, we believe that more work is needed on metrics for interdependent privacy, combinations of metrics, and evaluations of the quality of metrics.

In summary, we systematize the literature on privacy measurement. Our survey can thus serve as a reference guide for privacy metrics and as a framework that enables privacy researchers to make informed decisions on which metrics to choose in a particular setting. This will contribute to the advancement of PETs and privacy protection in general.

2 Conditions for Quality of Metrics

There is no general consensus which conditions high-quality privacy metrics have to fulfill. In the mathematical sense, a metric is a measure for the distance between two elements of a set, i.e., a function that maps each combination of two set members onto the set of real numbers: $d : X \times X \to \mathbb{R}$. Metrics have to fulfill four conditions, namely non-negativity $d(x, y) \geq 0$, identity of indiscernibles $d(x, y) = 0$ iff $x = y$, symmetry $d(x, y) = d(y, x)$, and the triangle inequality $d(x, z) \leq d(x, y) + d(y, z)$. Many of the metrics discussed in this survey are not metrics in the mathematical sense, as they do not fulfill all four conditions. Nevertheless, we will use the term 'privacy metric' to remain consistent with the literature.

In addition to the mathematical definition, many authors have proposed other conditions that are more specific to the measurement of privacy. Alexander and Smith [2003] require that privacy metrics are understandable by mathematically inclined laypeople, are orthogonal to cost and utility metrics, and give bounds on how effectively the adversary can succeed in identifying individuals. Andersson and Lundin [2008] require that privacy metrics are based on probabilities and have well defined and intuitive endpoints. In addition, anonymity should be rated higher the more uniform the probability distribution, and the more users there are in the anonymity set. Finally, they require that the metric's value domain should be ordered, not too coarse, and contain well-defined elements. Bertino et al. [2008] require that privacy metrics indicate the privacy level, the portion of sensitive data that is not hidden, and the data quality after application of the PET. Shokri et al. [2011] require that privacy metrics consider three aspects of the adversary's success: accuracy, uncertainty, and correctness. Syverson [2013] requires that privacy metrics reflect how difficult it is for an adversary to succeed, that they do not depend on variables that cannot be determined or predicted, and that they reflect the resources needed for successful attacks on privacy instead of relying on cardinalities or probabilities. In an earlier publication, we required that privacy metrics should be monotonous with increasing adversary strength [Wagner 2015].

3 Privacy Domains

Privacy domains are areas where privacy enhancing technologies (PETs) can be applied. With the increasing use of information technology, PETs are being researched in a growing number of domains. Here, we describe six domains to provide context and examples for the remainder of the paper.

3.1 Communication Systems

The main privacy challenge in communication systems is anonymous communication, which aims to hide which (or even that) two users communicated, not just the contents of their communication. Maintaining the confidentiality of communication contents is an orthogonal problem that can be solved via public-key encryption [Chaum 1988]. Adversaries typically try to identify either the sender of a message, its receiver, or sender-receiver relationships. Metrics for communication systems have been previously reviewed by Kelly et al. [2008].

3.2 Databases

There are two typical scenarios in the database domain: in the interactive setting, users issue queries to a database; in the non-interactive setting, a sanitized database is released to the public. In both scenarios, adversaries attempt to identify individuals in the database and reveal sensitive attributes, for example the health information contained in a patient record. Surveys that review metrics for this domain include Fung et al. [2010] (privacy preserving data publishing), Bertino et al. [2008] (data mining), and Kelly et al. [2008] (databases).

3.3 Location-based Services

Location-based services provide context-aware services to mobile users, such as information about nearby points of interest. Adversaries with access to location information can infer sensitive attributes like home and work locations, and create movement profiles that can be sold or used for marketing purposes. Metrics for location privacy are discussed by Shokri et al. [2010a] and Krumm [2009]. In previous work, we reviewed metrics for vehicular networks [Wagner and Eckhoff 2014].

3.4 Smart Metering

Smart meters record fine-grained electricity consumption data in a user's home and send this data to the energy provider. The energy provider can use this data for billing and network optimization, but can also act as an adversary who infers behavioral profiles above and beyond the stated purpose. Metrics and mechanisms for smart metering are reviewed by Zeadally et al. [2013].

3.5 Social Networks

Social networks allow users to share updates about their daily lives. Adversaries in this domain try to identify users in anonymized social graphs, or infer sensitive attributes from private profiles. Yang et al. [2012] survey privacy risks in social networks.

3.6 Genome Privacy

Advances in whole genome sequencing have raised new questions regarding the privacy of a person's genome. The genome uniquely identifies an individual, and at the same time reveals highly sensitive information, like susceptibility to diseases. An adversary with access to genomic data could engage in genetic discrimination (e.g., denial of insurance) or blackmail (e.g., planting fake evidence at crime scenes). In previous work, we reviewed privacy metrics for genomics [Wagner 2015].

4 Characteristics of Privacy Metrics

Despite their diversity, privacy metrics share common characteristics. Here, we describe four characteristics that can classify privacy metrics and can thus serve as an initial guideline for choosing privacy metrics for specific scenarios (we give detailed recommendations in Section 6).

4.1 Adversary Models

Many privacy metrics consider some kind of adversary because a stronger adversary leads to lower privacy. As a result, two PETs can only be directly compared if they use the same adversary model.

For example, entropy-based metrics assign probabilities to members of the anonymity set; these probabilities reflect the knowledge and strength of an adversary. A weaker adversary typically results in higher values of entropy, so a PET evaluated with a weak adversary model can appear to offer better privacy protection. Metrics that do not account for any type of adversary circumvent this problem, for example, by measuring certain properties of data. This implicitly assumes an adversary model where every attack on the system will only rely on the chosen properties, but attacks that exploit other characteristics of the data can disclose sensitive information.

The literature reflects the importance of adversary models by considering adversaries with diverse characteristics. We substantially extend the taxonomy of adversary types described by Diaz et al. [2003], and classify adversaries as follows:

4.1.1 Local–Global

Local adversaries can only act on a restricted part of the system, for example a geographical location or a subset of nodes. Global adversaries have access to the entire system.

4.1.2 Active–Passive

Active adversaries can interfere with the system by adding, removing or modifying information or communication. Passive adversaries can only read and observe.

4.1.3 Internal–External

Internal adversaries are part of the system, for example servers providing location-based services, energy providers in smart metering, or third parties controlling nodes in the system. External adversaries are not part of the system, but able to attack it, e.g., via shared communication links or publicly available data.

4.1.4 Static–Adaptive

Static adversaries choose which strategy and resources to use prior to an attack, and stick to their choice irrespective of how the attack progresses. Adaptive adversaries can adapt their attack while it is ongoing, e.g., by learning system parameters through observation.

4.1.5 Prior Knowledge

Some adversaries have additional knowledge about the system, such as general domain-specific knowledge – knowledge about the world – or scenario-specific knowledge, for example in the form of a prior probability distribution.

4.1.6 Resources

Adversaries can also be classified according to the resources available to them, especially their computational power. Efficient adversaries are restricted to probabilistic polynomial time (PPT) algorithms, while unbounded adversaries are not restricted to any computational model.

4.2 Data Sources

Data sources describe which data needs to be protected, and how the adversary is assumed to gain access to the data. We indicate the primary data sources for each metric in Tables 6.8 and 2 (column *Primary data source*).

4.2.1 Published Data

Published data refers to information that has been willingly and persistently made available to the public. This includes statistical databases as well as information individuals choose to disclose, e.g., on social networks. In both cases, adversaries attempt to identify anonymized individuals or reveal sensitive attributes.

4.2.2 Observable Data

Observable data is transient information that requires the adversary to be present in order to gain access to it. This category includes information that can be obtained by a passive adversary who can access data without compromising the underlying system. In communication systems, for example, adversaries overhear communications to identify message senders and receivers.

4.2.3 Re-purposed Data

Re-purposed data is used for a different purpose than the purpose for which it was initially acquired. Examples are service providers who obtain user information to offer location-based services, smart metering, or social networks, but then use this information for purposes other than providing the service. Having access to non-public user information (regardless of the users' privacy setting) allows for tailored advertising and other forms of marketing or monetization.

4.2.4 All Other Data

All other data refers to information that was not made public, was not observable and that the adversary was not intended to have access to. This data is typically not anonymized or protected, and can be obtained using methods such as wiretapping, hacking into a system, blackmailing, or buying off the black market. Implications for users can be severe, including financial losses and publication of medical records or confidential communication. PETs are often not deployed as they can make it less convenient to work with the data.

4.3 Inputs for Computation of Metrics

Privacy metrics rely on different kinds of input data to compute privacy values. The availability of input data or appropriate assumptions determine whether a metric can be used in a specific scenario. We indicate which of the input categories each metric relies on in Tables 6.8 and 2 (column group *Inputs*).

4.3.1 Adversary's Estimate

The adversary's estimate is the result of the adversary's effort to breach privacy. It often takes the form of a posterior probability distribution.

4.3.2 Adversary's Resources

The resources available to the adversary can be given, for example, in terms of computational power, bandwidth, or physical nodes.

4.3.3 True Outcome

The true outcome, or ground truth, is often used to judge how good the adversary's estimate is. However, this information is not available to the adversary, so they cannot compute metrics using the true outcome.

4.3.4 Prior Knowledge

Prior knowledge describes concrete, scenario-specific knowledge that the adversary has. It usually takes the form of a prior probability distribution.

4.3.5 Parameters

Parameters configure privacy metrics. They describe threshold values, the sensitivity of attributes, which attributes are sensitive, or desired privacy levels.

4.4 Output Measures

The output of a privacy metric refers to the kind of property that a privacy metric measures. We introduce a taxonomy with eight output properties, each of which represents a different aspect of privacy. This is an important categorization because it shows that a single metric cannot capture the entire concept of privacy. A more complete estimate of privacy can only be obtained by using metrics from different output categories. Figure 1 gives an overview of the output measures and the metrics associated with each.

4.4.1 Uncertainty

Uncertainty metrics assume that high uncertainty in the adversary's estimate correlates with high privacy, because the adversary cannot base his guesses on information known with certainty. However, even guesses based on uncertain information can be correct, and thus individual users may suffer privacy losses even in scenarios with a highly uncertain adversary.

4.4.2 Information Gain or Loss

Metrics that measure information gain or loss quantify the amount of information gained by the adversary, or the amount of information lost by users. These metrics assume that both high information gain and high information loss correlate with low privacy.

4.4.3 Similarity or Diversity

Similarity/diversity metrics measure the similarity or diversity between two sets of data, for example between a private dataset and its public, sanitized counterpart. Low similarity and high diversity between the two datasets correlate with high privacy. Most metrics in this category can be computed without considering the adversary; privacy is evaluated solely based on the data itself.

4.4.4 Indistinguishability

Indistinguishability is a classic notion in the security community. Metrics based on indistinguishability analyze whether the adversary is able to distinguish between two outcomes of a privacy mechanism. Privacy is high if the adversary cannot distinguish between any pair of outcomes. Metrics in this category are usually binary; they indicate whether two outcomes are indistinguishable or not, but do not quantify the privacy levels in-between.

4.4.5 Adversary's Success Probability

Metrics using the adversary's success probability to quantify privacy indicate how likely it is for the adversary to succeed in any one attempt, or how often they would succeed in a large number of

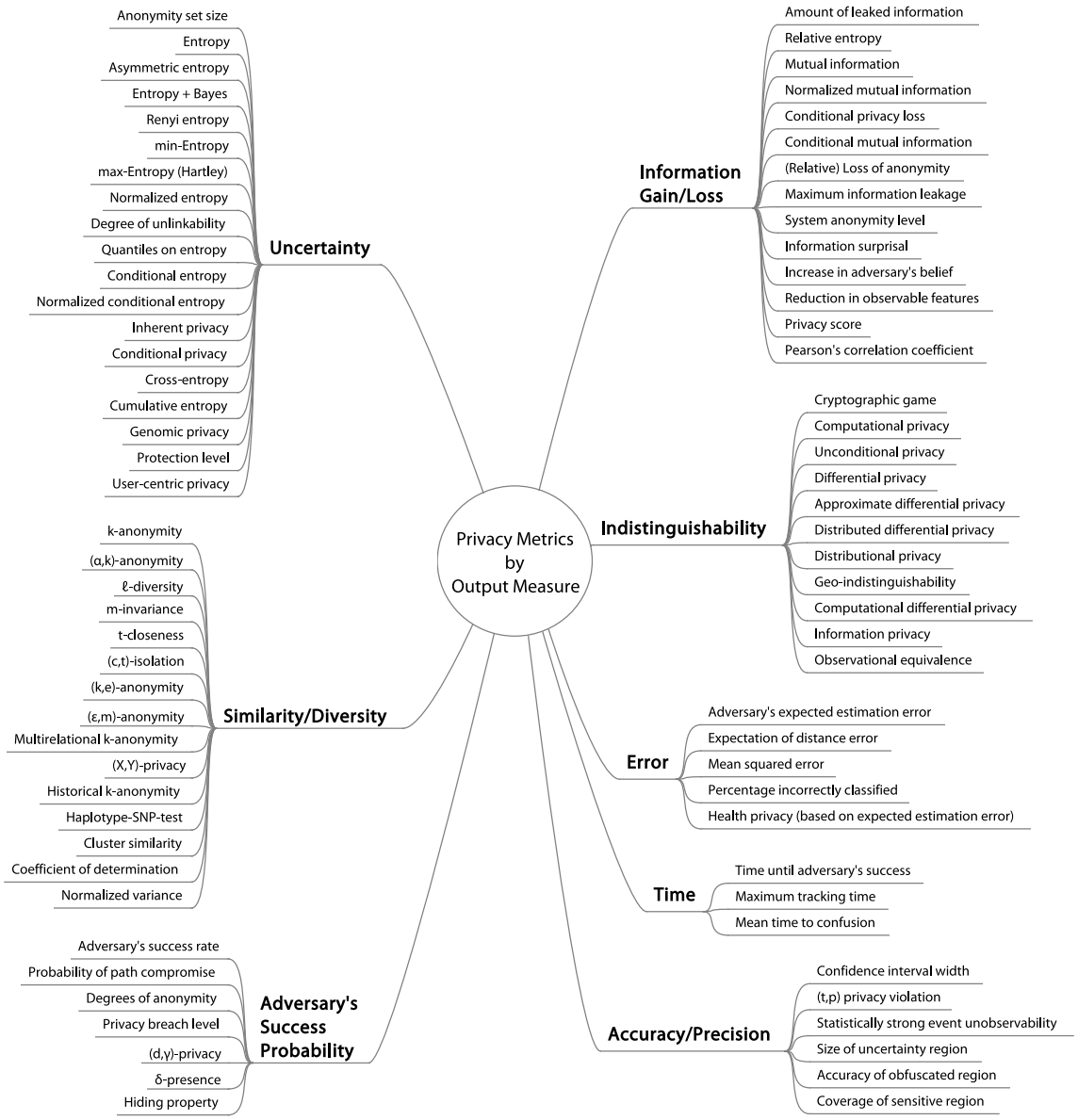

Figure 1: Taxonomy of privacy metrics, classified by output

attempts. Low success probabilities correlate with high privacy. While this assumption holds for an averaged population of users, an individual user may still suffer a loss of privacy even when the adversary's success probability is low.

4.4.6 Error

Error-based metrics measure how correct the adversary's estimate is, for example using the distance between the true outcome and the estimate. High correctness and small errors correlate with low privacy.

4.4.7 Time

Time-based metrics either measure the time until the adversary's success, or the time until the adversary's confusion. In the first case, metrics assume that the adversary will succeed eventually, and so a longer time correlates with higher privacy. In the second case, metrics assume that the privacy mechanism will eventually confuse the adversary, and so a shorter time correlates with higher privacy.

4.4.8 Accuracy or Precision

Accuracy metrics quantify how precise the adversary's estimate is without considering the estimate's correctness. A more precise estimate correlates with lower privacy.

5 Privacy Metrics

We now describe over eighty privacy metrics from the literature, grouped by the outputs they measure. Where possible, we point out their advantages or disadvantages and give examples for application scenarios. We also simplify and unify metric notation and point out similiarities or differences of related metrics.

At the end of the section, Tables 6.8 and 2 summarize how each metric can be classified according to the characteristics introduced in the previous section. The tables also provide information about value ranges, and an indication whether higher or lower values represent better privacy. We will refer to Tables 6.8 and 2 again in Section 6, when we give recommendations on how to select privacy metrics.

5.1 Uncertainty

Uncertainty metrics assume that an adversary who is uncertain of his estimate cannot breach privacy as effectively as one who is certain. Many uncertainty metrics build on entropy, an information theoretic notion to measure uncertainty [Shannon 1948].

5.1.1 Anonymity set size

The anonymity set size counts the number of users that could potentially be a targeted individual t [Chaum 1988; Kesdogan et al. 1998]. It can be seen as the size of the crowd into which the target t can blend.

$$priv_{\text{ASS}} \equiv |AS_t|$$

Instead of users, anonymity sets can also be applied to locations [Duckham and Kulik 2005], location pairs (e.g., home/work) [Golle and Partridge 2009], or radio frequency identification (RFID) devices [Heydt-Benjamin et al. 2006]. As a result of its simplicity, the anonymity set size is widely used in the literature.

The main criticism of the anonymity set size is that it only depends on the number of users in the system. This means that it does not take into account prior knowledge, information the adversary has gathered by observing the system, or how likely each member of the anonymity set is to be the target [Serjantov and Danezis 2002; Diaz et al. 2003]. However, it can be argued that the size of the anonymity set is useful in combination with other metrics such as normalized entropy (Section 5.1.4) [Steinbrecher and Köpsell 2003].

5.1.2 Entropy

Shannon entropy is the basis for many other metrics. In general, entropy measures the uncertainty associated with predicting the value of a random variable. As a privacy measure, it can be interpreted as the effective size of the anonymity set, or as the number of bits of additional information the adversary needs to identify a user [Serjantov and Danezis 2002]. The random variable X indicates the adversary's estimated probabilities for each member of the anonymity set.

$$priv_{\text{ENT}} \equiv H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

When the adversary has access to prior information about the distribution of the random variable X, the point w where entropy is maximum can differ for each element. For example, in genomics, information about the population-wide average probabilities of genomic variations are readily available. In this case, asymmetric entropy can be used instead of entropy to account for this prior information [Ayday et al. 2013b].

$$priv_{\text{AE}} \equiv \sum \frac{p_i(1-p_i)}{(-2w_i+1)p_i + w_i^2}$$

Entropy has also been used in cases where privacy is measured at more than one point in time, for example in location privacy, where the adversary tracks users during a period of time. In this case, entropy is computed at every point in time, and the underlying probabilities are updated after

each timestep using Bayesian belief tables [Ma et al. 2010]. After the first timestep, this accounts for the prior knowledge the adversary has acquired during previous timesteps.

Many papers argue against the use of entropy as a privacy metric. Entropy is strongly influenced by outlier values, i.e., users in the anonymity set that are very unlikely to be the target [Clauß and Schiffner 2006]. Even if an adversary is able to identify a target with high probability, the remaining low probability members of the anonymity set can still lead to high values of entropy [Tóth et al. 2004]. It is easy to construct different probability distributions that yield the same entropy value, for example a uniform distribution over 20 users, and an almost uniform distribution over 101 users where one user has a probability of $\frac{1}{2}$ [Tóth et al. 2004; Murdoch 2013]. This makes it difficult to compare different systems.

In the case of location privacy, entropy measures how well an adversary can disclose the position of a user. However, if two positions are very close to each other, locations may be revealed despite high entropy [Hoh and Gruteser 2005].

Although entropy has an intuitive interpretation as the number of additional bits of information the adversary needs, it can be argued that the absolute value of entropy does not convey much meaning [Hamel et al. 2011]. Entropy gives an indication of the adversary's uncertainty, but does not state how correct or accurate the adversary's estimates are [Shokri et al. 2011], or how many resources the adversary has to expend to succeed [Syverson 2013; Murdoch and Watson 2008].

5.1.3 Rényi Entropy

Rényi entropy is a generalization of Shannon entropy that also quantifies the uncertainty in a random variable. It uses an additional parameter α, and Shannon entropy is the special case with $\alpha \to 1$.

$$priv_{\mathrm{RE}} \equiv H_\alpha(P) = \frac{1}{1-\alpha} \log_2 \sum p_i^\alpha$$

Hartley entropy or max-entropy is the special case with $\alpha = 0$. It depends only on the number of users and is therefore a best-case scenario because it represents the ideal privacy situation for a user. Min-entropy is the special case with $\alpha = \infty$ which is a worst-case scenario because it only depends on the user for whom the adversary has the highest probability [Clauß and Schiffner 2006].

$$priv_{\mathrm{MXE}} \equiv H_0(X) = \log_2 |X|$$
$$priv_{\mathrm{MNE}} \equiv H_\infty(X) = -\log_2 \max_i p_i$$

5.1.4 Normalized Entropy (Degree of Anonymity)

Because the value range of entropy depends on the number of elements in the anonymity set, the absolute value can not always be used to compare entropy values. This is why entropy is frequently normalized using Hartley entropy (i.e., the maximum value entropy takes when all elements in the anonymity set are equally likely). Normalized entropy can be interpreted as the amount of information the system is leaking [Diaz et al. 2003].

$$priv_{\mathrm{NE}} \equiv \frac{H(X)}{H_0(X)}$$

5.1.5 Degree of Unlinkability

The degree of unlinkability measures the adversary's uncertainty about which items are related, for example which users are related by anonymous communication. The metric is based on entropy computed over the set of partitions Π of users U. Each partition π defines an equivalence class of related users, and the adversary aims to find the true partition τ [Steinbrecher and Köpsell 2003].

$$priv_{\mathrm{DUE}} \equiv H_\emptyset(U) = -\sum_{\pi \in \Pi} Pr(\pi = \tau) \log_2(Pr(\pi = \tau))$$

The degree of unlinkability can be extended to account for prior knowledge of an adversary by computing the ratio of the above degree of unlinkability for an adversary with (H_P) and without ($H_\emptyset$) prior knowledge [Franz et al. 2007].

$$priv_{\mathrm{DUP}} \equiv \frac{H_P(U)}{H_\emptyset(U)}$$

5.1.6 Quantiles on Entropy

Quantiles on entropy compute the entropy of a chosen percentile of the source alphabet, for example allowing to state that $c\%$ of all users "are at least as anonymous as the measurement states" [Clauß and Schiffner 2006].

$$priv_{\text{QE}} \equiv H(X), \text{ where } \forall x \in X : p(x) \geq c$$

5.1.7 Conditional Entropy

The conditional entropy, or equivocation, of a random variable X, given a random variable Y, measures how much information is needed to describe X if the value of Y is known. However, care must be taken to distinguish conditional entropy from the entropy of a conditional probability distribution [Diaz et al. 2007].

$$priv_{\text{COE}} \equiv H(X|Y) = -\sum_{x \in X, y \in Y} p(y,x) \log_2 p(x|y)$$

Normalized conditional entropy uses the entropy of X (because entropy is the maximum of conditional entropy) to normalize conditional entropy [Lai et al. 2011].

$$priv_{\text{NCE}} \equiv \frac{H(X|Y)}{H(X)}$$

5.1.8 Inherent Privacy

Inherent privacy (also called scaled anonymity set size) is derived from entropy and describes the privacy inherent in the random variable X as the number of possible outcomes given the expected amount of binary questions the adversary needs to answer [Agrawal and Aggarwal 2001; Andersson and Lundin 2008].

$$priv_{\text{IP}} \equiv 2^{H(X)}$$

In a similar way, conditional privacy is based on conditional entropy and measures the privacy inherent in a random variable X, given random variable Y [Agrawal and Aggarwal 2001].

$$priv_{\text{CP}} \equiv 2^{H(X|Y)}$$

5.1.9 Cross-entropy / Likelihood

In data clustering, cross entropy measures the uncertainty in predicting the original dataset from the clustered model [Merugu and Ghosh 2003]. Generally, cross entropy measures the amount of information needed to identify an object in the data set if the original data are coded in terms of the model's distribution q, rather than their true distribution p. Cross entropy is derived from entropy (Section 5.1.2) and the relative entropy D_{KL} (Section 5.2.2).

$$priv_{\text{CE}} \equiv H(p) + D_{\text{KL}}(p||q)$$

5.1.10 Cumulative Entropy

In location privacy, cumulative entropy measures how much entropy can be gathered on a route through a series of independent mix zones. A mix zone M is an area where several vehicles are close to each other at the same time, such that the adversary cannot distinguish the vehicles as they leave the mix zone in different directions. Cumulative entropy adds up the entropy gathered in each mix zone m on a vehicle's path [Freudiger et al. 2007].

$$priv_{\text{CUE}} \equiv \sum_{m=1}^{M} H_m(X)$$

5.1.11 Genomic Privacy

Genomic privacy measures the probability that a specific variation (a so-called single nucleotide polymorphism, or single nucleotide polymorphism (SNP)) occurs in a person's genome, and weights this probability with a rating W_i of the SNP's severity, which indicates, for example, how much this SNP contributes to a disease [Ayday et al. 2013a].

$$priv_{\text{GP}} \equiv -\sum_{i \in \text{SNP}} \log_2(Pr(\text{SNP}_i = 1)) \cdot W_i$$

5.1.12 Protection Level

The protection level is a metric from location privacy which is based on the popularity of a region $P(R)$. Popularity is defined as the inherent privacy (Section 5.1.8) computed over the frequencies of location samples from users in this region. A user in the system can specify a public reference region to define how private they want to be. The protection level is then the ratio of the average popularity of the regions along his trajectory and the popularity of the reference region, normalized with the set of users common to all regions T [Xu and Cai 2009].

$$priv_{\text{PL}} \equiv \frac{\sum_i P_T(R_i)}{|T|P_T(R)}, \text{ where popularity } P_T(R) = 2^{H(R)}$$

5.1.13 User-centric Privacy

In location privacy, user-centric privacy assumes that privacy decreases over time, and is increased when privacy mechanisms are applied at regular points in time T_i^ℓ. The decrease of privacy is caused by the adversary collecting additional information, and modeled with a privacy loss function $\beta_i(t, T_i^\ell)$. User-centric privacy thus combines a base privacy metric such as entropy $H(X, T_i^\ell)$ with a linear decay of privacy over time [Freudiger et al. 2009]. Each user specifies a sensitivity parameter λ_i, which governs how quickly their privacy decays.

$$priv_{\text{UCP}} \equiv H(X, T_i^\ell) - \beta_i(t, T_i^\ell), \text{ where } T_i^f = \frac{H(X, T_i^\ell)}{\lambda_i} + T_i^\ell \text{ and}$$

$$\beta_i(t, T_i^\ell) = \begin{cases} \lambda_i \cdot (t - T_i^\ell) \text{ for } T_i^\ell \leq t < T_i^f \\ H(X, T_i^\ell) \text{ for } T_i^f \leq t \end{cases}$$

5.2 Information Gain or Loss

Information gain metrics measure the amount of information an adversary can gain, assuming that privacy is higher the less information an adversary can gain. Similar to uncertainty metrics, many information gain metrics are based on information theory. However, information gain metrics explicitly consider the amount of prior information.

5.2.1 Amount of Leaked Information

This metric counts the number of information items L disclosed by a system, e.g., the number of compromised users [Backstrom et al. 2007] or the number of leaked DNA base pairs [Wang et al. 2009]. However, this metric does not indicate the severity of a leak because it does not account for the sensitivity of the leaked information.

$$priv_{\text{ALI}} \equiv |L|$$

5.2.2 Relative Entropy

Relative entropy (also called Kullback-Leibler divergence D_{KL}) measures the distance between two probability distributions p and q. As a privacy metric, the two distributions usually represent the true distribution p and the adversary's estimate q, and relative entropy represents the amount of probabilistic information revealed to the adversary [Deng et al. 2007].

$$priv_{\text{RLE}} \equiv D_{\text{KL}}(p||q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)}$$

5.2.3 Mutual Information

Mutual information quantifies how much information is shared between two random variables. It can be computed as the difference between entropy (Section 5.1.2) and conditional entropy (Section 5.1.7). Mutual information measures the amount of information leaked from a privacy mechanism [Lin et al. 2002].

$$priv_{\text{MI}} \equiv I(X;Y) = H(X) - H(X|Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Mutual information between X and Y can be normalized using the entropy of X to allow comparisons between scenarios. This can be interpreted as the degree of dependence between hidden data X and observed data Y [Humbert et al. 2013].

$$priv_{\text{NMI}} \equiv 1 - \frac{I(X;Y)}{H(X)}$$

5.2.4 Conditional Privacy Loss

Another way of normalizing mutual information is the conditional privacy loss, which measures the fraction of privacy of X which is lost by revealing Y [Agrawal and Aggarwal 2001].

$$priv_{\text{CPL}} \equiv 1 - 2^{-I(X;Y)}$$

5.2.5 Conditional Mutual Information

Mutual information can also be applied when the adversary has access to prior knowledge. Conditional mutual information measures the amount of information about X that can be learned by observing Y, given prior knowledge Z. It measures the correlation between X and Y given Z [Coble 2008].

$$priv_{\text{CMI}} \equiv I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

5.2.6 (Relative) Loss of Anonymity

Loss of anonymity describes the amount of information that can be learned about a set of anonymous events X, given a set of observed events Y, for the least private distribution of X [Chatzikokolakis et al. 2007]. The metric is based on mutual information (Section 5.2.3), and finds the least private distribution by taking a maximum over all possible distributions of X.

$$priv_{\text{LA}} \equiv max_{p(x)} I(X;Y)$$

Relative loss of anonymity extends loss of anonymity by taking into account that the adversary has access to certain revealed information Z. Instead of mutual information, this metric is based on conditional mutual information (Section 5.2.5).

$$priv_{\text{RLA}} \equiv max_{p(x)} I(X;Y|Z)$$

5.2.7 Maximum information leakage

Maximum information leakage modifies mutual information to consider only a single realization of the random variable Y. It quantifies the maximum amount of information about private events or data X that can be gained by an adversary observing a single output y [du Pin Calmon and Fawaz 2012].

$$priv_{\text{MIL}} \equiv max_{y \in Y} H(X) - H(X|Y=y)$$

5.2.8 System Anonymity Level

In anonymous communication, the system's anonymity level describes the amount of additional information needed to reveal all sender-receiver relationships. Sender-receiver relationships are described in the adjacency matrix A, and the system's anonymity level corresponds to finding the equivalence class containing the correct perfect matching between senders and receivers. The metric is based on computing entropy using the frequency of perfect matchings C_p in the permanent of the adjacency matrix $per(A)$, normalized with the number of users U [Gierlichs et al. 2008].

$$priv_{\text{SAL}} \equiv \begin{cases} 0, & \text{if } |U| = 1 \\ \frac{1}{\log(|U|!)} H(\frac{C_p}{per(A)}), & \text{if } |U| > 1 \end{cases}$$

5.2.9 Information Surprisal

Information surprisal is a measure of self-information. It quantifies how much information is contained in a specific outcome x of a random variable X. In social networks, information surprisal measures how much information the adversary gains if he gets to know the value of one specific attribute in a user profile [Chen et al. 2013].

$$priv_{\text{IS}} \equiv -\log_2 P(X = x)$$

5.2.10 Increase in Adversary's Belief

The increase in adversary's belief measures the difference between the adversary's prior and posterior probabilities (e.g., of identifying an individual x in a set of users X). Privacy is breached if this difference is greater than the privacy parameter δ [Narayanan and Shmatikov 2009].

$$priv_{\text{IAB}} \equiv \delta, \text{ where } Pr_{\text{post}}(X = x) - Pr_{\text{prior}}(X = x) > \delta$$

5.2.11 Reduction in Observable Features

In smart metering, load hiding algorithms try to hide load transitions from the energy provider. The reduction in observable features measures how many transitions are hidden successfully [McLaughlin et al. 2011]. The transitions form a time-series D with length w, which is condensed to a single value, the feature mass F, for example the number of non-zero energy transitions d_i. The metric then relates the feature masses with (D_P) and without (D_T) privacy protection. However, a reduction in observable features does not necessarily mean that sensitive information cannot be inferred.

$$priv_{\text{ROF}} \equiv \frac{F(w, D_P)}{F(w, D_T)}, \text{ where } FM(w, D) = \sum_{i=0}^{w} (d_i \neq 0)$$

5.2.12 Privacy Score

The privacy score in a social network indicates a user j's potential privacy risk. It increases with the sensitivity β_i of information items i and their visibility $V(i, j)$, e.g., the number of users knowing about each item [Liu and Terzi 2010].

$$priv_{\text{PS}} \equiv \sum_{i=1}^{n} \beta_i \times V(i, j)$$

5.2.13 Pearson's Correlation Coefficient

In statistics, Pearson's correlation coefficient measures the degree of linear dependence between two random variables. It is computed as the covariance between X and Y, normalized with the standard deviations σ_X and σ_Y. In smart metering, this can be used to measure the correlation between original and obfuscated load data [Kim et al. 2011].

$$priv_{\text{PCC}} \equiv \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

5.3 Similarity or Diversity

Similarity and diversity metrics measure similarity and diversity properties of observable or published data. These metrics are usually independent of the adversary and derive the privacy level solely from the similarity or diversity of data.

5.3.1 k-Anonymity

k-Anonymity is conceptually similar to the size of the anonymity set (Section 5.1.1), but does not consider the adversary. It was originally proposed for statistical databases, where k indicates the number of rows in a database table D that are indistinguishable with respect to their quasi-identifiers q [Sweeney 2002]. Quasi-identifiers by themselves do not identify users, but can do so when correlated with other data. The sets of indistinguishable rows partition a table into equivalence classes with a minimum size of k.

$$priv_{\text{KA}} \equiv k, \text{ where } \forall(c = \text{ combination of values of } q) : |D[c]| \geq k$$

Several algorithms exist that transform a given database to make it k-anonymous by suppression or generalization [Samarati and Sweeney 1998]. However, studies have shown k-anonymity to be insufficient, especially for high-dimensional data [Aggarwal 2005] and against correlation with other data sets [Machanavajjhala et al. 2007], because it fails to protect against attribute disclosure [Xiao and Tao 2006]. In addition, k-anonymous data releases do not offer protection across multiple releases of the same data set [Xiao and Tao 2007], or when sensitive data, such as location data, are semantically close [Shokri et al. 2010b]. Despite this criticism, k-anonymity is still widely used today, and is routinely applied to new privacy domains.

5.3.2 (α, k)-Anonymity / Privacy Templates

This metric extends k-anonymity with the additional requirement that in any equivalence class E (defined by combinations of quasi-identifier values q), the frequency of a sensitive value s has to be less than α [Wong et al. 2006; Wang et al. 2007]. As a result, no single sensitive attribute can be dominant in an equivalence class.

$$priv_{\text{AK}} \equiv (\alpha, k), \text{ where } \forall(c = \text{ combination of values of } q) : |D[c]| \geq k \wedge \forall E : \frac{|(E, s)|}{|E|} \leq \alpha$$

However, it has been shown that attribute linkage can occur even when the frequency of s is less than α [Fung et al. 2010].

5.3.3 ℓ-Diversity

The ℓ-diversity principle modifies k-anonymity to bound the diversity of published sensitive information. It states that for every set of quasi-identifier tuples in the same equivalence class E, the ℓ most frequent values of the sensitive attribute s must have roughly the same frequencies $p_{E,s}$ [Machanavajjhala et al. 2007]. This general principle can be instantiated in different ways. In an instantiation based on entropy, for example, similar frequencies are indicated by a high entropy of the sensitive attribute frequencies.

$$priv_{\text{LE}} \equiv \ell, \text{ where } \forall E : -\sum_{s \in S} p_{(E,s)} \log(p_{(E,s)}) \geq \log(\ell)$$

In an instantiation based on recursion, the most frequent value s_1 must occur less often than all other values s_i combined, within a multiplicative factor c.

$$priv_{\text{LR}} \equiv \ell, \text{ where } \forall E : s_1 < c(s_\ell + s_{\ell+1} + ... + s_n)$$

Although ℓ-diversity is an improvement to k-anonymity, it has been shown to offer insufficient protection against some attacks. In particular, it does not protect privacy when multiple releases of statistical data are available [Xiao and Tao 2007], when the distribution of sensitive values is skewed [Li et al. 2007], or when sensitive attributes are semantically similar [Li et al. 2007]. In addition, the adversary may be able to reconstruct sensitive attributes if he knows the algorithm used for data sanitization [Zhang et al. 2007a]. Like k-anonymity, ℓ-diversity does not sufficiently protect sensitive attributes that are numerical rather than categorical [Zhang et al. 2007b].

5.3.4 m-Invariance

m-Invariance modifies k-anonymity to allow for multiple releases of the same data set that may contain added, modified, or deleted rows. m-Invariance states that every equivalence class E must have at least m rows, and the values for sensitive attributes s must all be different [Xiao and Tao 2007]. In addition, for any row the set of distinct sensitive values in its equivalence class must be the same in every release.

$$priv_{\text{MI}} \equiv m, \text{ where } \forall E : |E| \geq m \ \wedge \forall s_i, s_j : s_i \neq s_j \ \wedge$$
$$\forall s_i : \text{ distinct } s \text{ must be the same in all releases}$$

5.3.5 t-Closeness

t-closeness modifies k-anonymity to bound the distribution of sensitive values. It states that the distribution of sensitive values S_i in any equivalence class must be close to their distribution in the overall table S. In particular, the distance between distributions $d(S, S_i)$, measured using the Earth Mover Distance metric, must be smaller than a threshold t [Li et al. 2007].

$$priv_{\text{TC}} \equiv t, \text{ where } \forall (c = \text{ combination of values of } q) : d(S, S_i) \leq t$$

5.3.6 (c,t)-Isolation

(c,t)-Isolation extends k-anonymity to consider an adversary. The metric measures how well an adversary can isolate points in a database D [Chawla et al. 2005]. The difference between the adversary's guess y_{Adv} and the target point y_D is given by δ_y. A target point y_D is (c,t)-isolated if a sphere S with radius $c\delta_y$ around the adversary's guess includes less than t other points. c can be seen as isolation parameter, determining the size of the ball, whereas t is a privacy threshold.

$$priv_{\text{CT}} \equiv (c,t), \text{ where } |S(y_{Adv}, c\delta_y) \cap D| < t \text{ and } \delta_y = \|y_{Adv} - y_D\|$$

5.3.7 (k,e)-Anonymity

(k,e)-anonymity modifies k-anonymity to apply to numerical instead of categorical attributes. In addition to the k-anonymity requirement, it requires that the range of sensitive attributes in any equivalence class E must be greater than e [Zhang et al. 2007b].

$$priv_{\text{KE}} \equiv (k,e), \text{ where } \forall E : |E| \geq k \wedge range(E) > e$$

However, (k,e)-anonymity does not take into account how values within the range e are distributed, which can lead to attribute disclosure via a proximity attack [Fung et al. 2010; Li et al. 2008].

5.3.8 (ϵ, m)-Anonymity

(ϵ, m)-anonymity also modifies k-anonymity to apply to numerical attributes. It addresses the proximity attack against (k, e)-anonymity by bounding the probability of inferring the value of a sensitive attribute to at most $1/m$. To achieve this bound, (ϵ, m)-anonymity demands that in each equivalence class E and for each sensitive value x in E, at most $1/m$ of the rows can have a sensitive value s that is similar to x [Li et al. 2008].

$$priv_{\text{EM}} \equiv (\epsilon, m), \text{ where } \forall E, \forall x \in S : \frac{|x \in [s-\epsilon, s+\epsilon]|}{|E|} \leq \frac{1}{m}$$

5.3.9 Multirelational k-Anonymity

Multirelational k-anonymity modifies k-anonymity to apply to the record owner level instead of the record level, thus extending it to tables in a relational database [Nergiz et al. 2009]. To do this, multirelational k-Anonymity joins the database table identifying the record owners D_{pers} with all tables containing database records D_i, and then applies k-anonymity to the result of the join J.

$$priv_{\text{MK}} \equiv k, \text{ where } J = D_{pers} \bowtie D_1 \bowtie \cdots \bowtie D_n \text{ and}$$
$$\forall(c = \text{ combination of values of } q) : |J[c]| \geq k \text{ and } \forall(d_p \in D_{pers}) : |J[d_p]| \geq k$$

5.3.10 (X, Y)-Privacy

(X, Y)-privacy modifies k-anonymity to cope with sequential data releases by limiting the amount of linkage between two data sets [Wang and Fung 2006]. X and Y denote groups of database columns with quasi-identifiers and sensitive properties, respectively. (X, Y)-privacy limits the confidence with which sensitive values in Y can be inferred by requiring that the for any $x \in X$ and $y \in Y$, the percentage of records containing x and y, among those containing x, be less than k. Applied to sequential data releases, (X, Y)-privacy can be used to specify generalization rules that make sequential releases (X, Y)-private.

$$priv_{\text{XY}} \equiv k, \text{ where } \max_{y \in Y}\{\max_{x \in X}\{\frac{|D[y,x]|}{|D[x]|}\}\} \leq k, \text{ where } 0 < k \leq 1$$

5.3.11 Historical k-Anonymity

Historical k-anonymity extends k-anonymity to location privacy by requiring that the adversary can only link a user request to k or more users [Bettini et al. 2005]. To formalize this requirement, a user's personal history of locations (PHL) is defined to be time-location consistent with a set of requests R if the time and location of each request in R can be matched to a time and location in the PHL. Historical k-anonymity is satisfied if a user's set of requests R_u is location-time consistent with the PHLs of $k-1$ other users U.

$$priv_{\text{HKA}} \equiv k, \text{ where } \forall u, v \in U : |\text{PHL}_v \text{ is location-time consistent with } R_u| \geq k$$

5.3.12 Haplotype-SNP-Test

In genomics, the haplotype-SNP-test determines conditions when data about genomes, such as aggregate data or test statistics (p-values), can safely be published. The test is based on the number of genomic variations L in a study and the number of participants N. The threshold α limits the usefulness of published test statistics to the adversary [Zhou et al. 2011].

$$priv_{\text{HS}} \equiv \begin{cases} \frac{2(|N|-1)}{\log(|N|+1)} > |L|, & \text{condition for aggregate data} \\ \frac{2(|N|-1)}{\log(|N|+1)-1+\alpha} > |L|, & \text{condition for test statistics} \end{cases}$$

5.3.13 Cluster Similarity

In smart metering, the time series of differences in load measurements, so-called transitions, can be obfuscated by a load hiding algorithm. Cluster similarity measures how close original time series D_T and obfuscated time series D_P are [Kalogridis et al. 2010]. To compute cluster similarity, original and obfuscated time series are clustered into clusters C_T and C_P, respectively. The ratio of incorrectly classified transitions measures how effectively the original values have been hidden.

$$priv_{\text{CS}} \equiv \frac{|\forall i : C_{Ti} \cap C_{Pi}|}{|D_T|}$$

5.3.14 Coefficient of Determination R^2

The coefficient of determination R^2 measures how much variability in data is accounted for by a model for the data. In smart metering, the model is a linear regression fitted to obfuscated load transitions $dp(t)$, resulting in predicted values $\hat{dp}(t)$ [Kalogridis et al. 2010]. The coefficient of determination compares the error sum of squares SS_E and the regression sum of squares SS_R.

$$priv_{\text{R2}} \equiv 1 - \frac{SS_E}{SS_R + SS_E}, \text{ where } SS_E = \sum_t (dp(t) - \hat{dp}(t))^2 \text{ and } SS_R = \sum_t (\hat{dp}(t) - \bar{dp(t)})$$

5.3.15 Normalized Variance

In privacy-preserving data publishing that uses data perturbation, normalized variance is derived from the statistical variance σ^2 and measures the dispersion between the original X and perturbed data Y [Oliveira and Zaïane 2003].

$$priv_{\text{VAR}} \equiv \frac{\sigma^2(X - Y)}{\sigma^2(X)}$$

5.4 Indistinguishability

Indistinguishability metrics indicate whether the adversary can distinguish between two items of interest (such as recipients of a message, or sensitive attributes in a database). Many of these metrics are associated with privacy mechanisms that provide formal privacy guarantees. Most metrics have a binary outcome indicating whether the guarantees are fulfilled or not.

5.4.1 Cryptographic Games

Cryptographic games can be used to prove security and privacy properties of cryptographic protocols. They consist of a challenge-response game in which the adversary selects the inputs for a protocol, and is given the output and two alternative outcomes s and s' after the protocol has been executed. The adversary then has to determine whether s or s' is the correct outcome. The adversary has an advantage if they can do this with a probability that is non-negligibly greater than $\frac{1}{2}$, that is, if their probability is better than a random guess [Juels and Weis 2009]. The security or privacy property under investigation holds if the adversary's advantage is smaller than a negligible function $\epsilon(k)$ (k is a security parameter).

$$priv_{\text{CG}} \equiv \begin{cases} 1 \text{ if } Pr[s_{Adv} = s] \leq \frac{1}{2} + \epsilon(k) \\ 0 \text{ otherwise} \end{cases}$$

The computational privacy metric uses a cryptographic game with negligible advantage, while unconditional privacy uses an advantage of zero [Hermans et al. 2011].

5.4.2 Differential Privacy

In statistical databases, differential privacy guarantees that any disclosure is equally likely (within a small multiplicative factor ϵ) regardless of whether or not an item is in the database [Dwork 2006]. This guarantee is usually achieved by adding a small amount of random noise to the results of database queries. Formally, differential privacy is defined using two data sets D_1 and D_2 that differ in a single row. A randomized function $\mathcal{K}$ operating on these data sets is ϵ-differentially private if for all sets of query responses S, the output random variables (query responses) for the two data sets differs by at most $exp(\epsilon)$.

$$priv_{\text{DP}} \equiv Pr[\mathcal{K}(D_1) \in S] \leq exp(\epsilon) \times Pr[\mathcal{K}(D_2) \in S]$$

In the interactive setting, differential privacy provides privacy guarantees if the allowed number of queries is limited [McSherry 2009] (each subsequent query reduces the strength of the privacy guarantee by adding its privacy parameter ϵ). In the non-interactive setting [Dwork et al. 2009], differential privacy provides guarantees only for a certain class of queries [Soria-Comas and Domingo-Ferrert 2013]. In addition, the choice of the parameter ϵ is difficult: values reported in the literature vary from 0.01 [Hsu et al. 2014] to 100 [Yu et al. 2014]. A no-free-lunch theorem shows that differential privacy's guarantees degrade in the case of correlated data, for example when nodes are added to a social network graph [Kifer and Machanavajjhala 2011].

5.4.3 Approximate Differential Privacy

Approximate differential privacy relaxes differential privacy by allowing an additional small additive constant δ [Dwork et al. 2006].

$$priv_{\text{ADP}} \equiv Pr[\mathcal{K}(D_1) \in S] \leq exp(\epsilon) \times Pr[\mathcal{K}(D_2) \in S] + \delta$$

5.4.4 Distributed Differential Privacy

Distributed differential privacy extends approximate differential privacy to a distributed setting so that the data aggregator can be untrusted and collude with a subset of the participants [Shi et al. 2011]. This extension can be useful in smart metering, where users may not trust the energy provider (who acts as data aggregator). In distributed differential privacy, the output random variables (query responses) are conditioned on the randomness r_K from compromised participants. This ensures that differential privacy is achieved using only randomness provided by honest participants.

$$priv_{\text{DDP}} \equiv Pr[\mathcal{K}(D_1) \in S|r_K] \leq exp(\epsilon) \times Pr[\mathcal{K}(D_2) \in S|r_K] + \delta$$

5.4.5 Distributional Privacy

Distributional privacy extends differential privacy to a setting in which the data sets themselves do not need to be protected, but instead the parameters governing the generation of data. In a smart metering scenario, these parameters can be user habits, behavioral patterns, or sets of appliances in a home [Jelasity and Birman 2014]. Distributional privacy assumes a distributed setting in which smart meters apply noise to their local data, limiting the energy provider to querying this distributed database. Formally, distributional privacy uses two parameter sets θ_1 and θ_2 which govern the creation of two data sets and differ in at most one element. A randomized function $\mathcal{K}$ is distributionally ϵ-differentially private if the series of query responses $\mathcal{K}_j$ for the two parameter sets differs by at most $exp(\epsilon)$.

$$priv_{\text{DSP}} \equiv Pr[\theta_1|\mathcal{K}_j] \leq Pr[\theta_2|\mathcal{K}_j] \cdot \exp(\epsilon)$$

5.4.6 Geo-Indistinguishability

Geo-indistinguishability extends differential privacy to location privacy scenarios. It states that any two locations x_1, x_2 within a given radius $d(\cdot\,,\cdot\,)$ produce observations with distributions that are similar within a small multiplicative factor ϵ [Andrés et al. 2013].

$$priv_{\text{GI}} \equiv d(\mathcal{K}(x_1), \mathcal{K}(x_2)) \leq \epsilon d(x_1, x_2)$$

5.4.7 Computational Differential Privacy

Computational differential privacy replaces the unrestricted adversary used in differential privacy with a computationally bounded adversary. By using a weaker adversary model, computationally differentially private mechanisms can give more accurate query responses. Informally, computational differential privacy requires that the outputs produced by the privacy mechanism "look" differentially private to every adversary. Depending on how "look" is formalized, the definitions of computational differential privacy can be different [Mironov et al. 2009]. For example, a definition based on indistinguishability replaces the unrestricted adversary with a computationally bounded adversary A_κ.

$$priv_{\text{IND\_CDP}} \equiv Pr[A_\kappa(\mathcal{K}_\kappa(D_1)) = 1] \leq exp(\epsilon_\kappa) \times Pr[A_\kappa(\mathcal{K}_\kappa(D_2)) = 1] + \text{negl}(\kappa)$$

A definition based on simulation requires that the outputs from randomized functions R_κ are computationally indistinguishable from the outputs from ϵ-differentially private mechanisms $\mathcal{K}_\kappa$.

$$priv_{\text{SIM\_CDP}} \equiv |Pr[A_\kappa(\mathcal{K}_\kappa(D)) = 1] - Pr[A_\kappa(R_\kappa(D)) = 1]| \leq \text{negl}(\kappa)$$

5.4.8 Information Privacy

Information privacy captures the notion that the prior and posterior probabilities of inferring sensitive data s do not change significantly, given query outputs u. ϵ-information privacy implies 2ϵ-differential privacy, but additionally bounds the maximum information leakage (Section 5.2.7) [du Pin Calmon and Fawaz 2012]. Formally, a privacy preserving mapping $p_{S|U}(\cdot\,|\,\cdot)$ provides ϵ-information privacy

if for all sensitive values s, the ratio of posterior probability $p_{S|U}(s|u)$ to prior probability $p_S(s)$ is very close to 1.

$$priv_{\text{IP}} \equiv exp(-\epsilon) \leq \frac{p_{S|U}(s|u)}{p_S(s)} \leq exp(\epsilon),\ \forall u \in U : p_U(u) > 0$$

In the context of wireless sensor networks, information privacy expresses that event sources cannot be observed by an adversary. Event source unobservability requires that for all possible observations of events in a system, the adversary's prior probability equals the posterior [Yang et al. 2008].

5.4.9 Observational Equivalence

Observational equivalence is a formal property that states that the adversary cannot distinguish between two situations [Hughes and Shmatikov 2004]. To use this metric, privacy protocols are modeled using a formal process calculus such as the applied π-calculus. Then, privacy properties can be defined and verified using observational equivalence, i.e., a privacy property is fulfilled if the observable outputs from protocol runs in two situations are equivalent. This has been used, e.g., in voting privacy [Delaune et al. 2009], mobile telephony [Arapinis et al. 2012] and webs of trust [Backes et al. 2010].

5.5 Adversary's Success Probability

Metrics based on the adversary's success probability can be seen as general-purpose metrics that subsume many other aspects of privacy. They depend strongly on the adversary model and on how exactly success is defined.

5.5.1 Adversary's Success Rate

This metric measures the probability that the adversary is successful, or the percentage of successes in a large number of attempts [Wright et al. 2003]. Depending on the application scenario, success can be defined in different ways: in databases, the adversary is successful when he can find a record r' that is similar to the target record r with a similarity threshold of θ and an error threshold of ω [Narayanan and Shmatikov 2008].

$$priv_{\text{SRD}} \equiv Pr[Sim(r, r') \geq \theta] \geq \omega$$

In communication systems, the adversary is successful when he can identify the sender of a message [Shmatikov 2002], or when he can compromise a communication path with a given amount of resources (e.g., number of nodes and bandwidth) [Murdoch and Watson 2008].

5.5.2 Degrees of Anonymity

Reiter and Rubin [1998] define six degrees of anonymity for communication systems, which depend on how likely the adversary's success is. 'Absolute privacy' states that there are no observable effects. 'Beyond suspicion' indicates that the sender is equally as likely as all other potential senders. 'Probable innocence' means that the sender is as likely as not to be the originator of a message. 'Possible innocence' states that there is a nontrivial probability that the sender is someone else. 'Exposed' indicates that the adversary's probability is above a threshold θ. Lastly, 'provably exposed' says that the adversary can prove who the sender is.

$$priv_{\text{DOA}} \equiv \begin{cases} \text{absolute privacy}, & \text{if } Pr = 0 \\ \text{beyond suspicion}, & \text{if } Pr = \min \\ \text{probable innocence}, & \text{if } Pr \leq 0.5 \\ \text{possible innocence}, & \text{if } Pr < \max \\ \text{exposed}, & \text{if } Pr \geq \theta \\ \text{provably exposed}, & \text{if } Pr = 1 \end{cases}$$

However, it has been noted that the degree of anonymity does not reflect the adversary's real probability of success, because it ignores the cardinality of the anonymity set [Murdoch 2013].

User-specified innocence [Chen and Pang 2012] merges two degrees of anonymity, probable and possible innocence, by introducing a parameter α that represents the probability of the most likely user in the anonymity set.

5.5.3 Privacy Breach Level

A privacy breach occurs if the posterior probability of a property, given its prior probability, is higher than the threshold ρ. In a data mining scenario where a server mines association rules between items based on their occurrence in user transactions, the privacy breach level uses the probability that an item a is contained in a transaction t_i, given that the item is part of an item set A, which is a subset of the randomized transaction t'_i [Evfimievski et al. 2004].

$$priv_{\text{PBL}} \equiv \rho, \text{ where } \exists a \in A \text{ so that } P[a \in t_i | A \subseteq t'_i] \geq \rho$$

The privacy breach level can also measure privacy in networking, where the metric refers to the conditional probability that node A generated a message with specific characteristics, given that node B received such a message [Seys and Preneel 2009].

An extension of the privacy breach level is d, γ-privacy, which introduces an additional bound d on the prior probability so that the posterior probability γ cannot drop by more than a factor of d/γ [Rastogi et al. 2007].

$$priv_{\text{DG}} \equiv \frac{d}{\gamma} \leq \frac{Pr_{\text{post}}(a|A)}{Pr_{\text{prior}}(a)}, \text{ where } Pr_{\text{prior}}[a] \leq d \text{ and } Pr_{\text{post}}[a|A] \leq \gamma$$

5.5.4 δ-Presence

In the database domain, δ-presence bounds the adversary's probability of inferring that an individual y is in a database D_T between δ_{min} and δ_{max}. The metric assumes that the adversary knows the published database D_P and has access to prior knowledge P (e.g., external tables) [Nergiz et al. 2007].

$$priv_{\text{DLP}} \equiv (\delta_{min}, \delta_{max}), \text{ where } \forall y \in P : \delta_{min} \leq Pr(y \in D_T | D_P) \leq \delta_{max}$$

However, this model assumes that the data publisher and the adversary have access to the same external tables. This assumption may not hold in practice [Fung et al. 2010].

5.5.5 Hiding Property

In communication systems, the source (or destination) hiding property measures the adversary's maximum probability for any user to be sender (or recipient) of a given message. The source (or destination) is assumed to be hidden if this probability is smaller than a threshold θ [Tóth et al. 2004].

$$priv_{\text{HP}} \equiv \theta, \text{ where } \forall_m \forall_u (P_{m,u} \leq \theta)$$

5.6 Error

Error-based metrics quantify the error an adversary makes in creating his estimate. Because information about the true outcome is needed to compute these metrics, they cannot be computed by the adversary.

5.6.1 Adversary's Expected Estimation Error

In location privacy, the adversary's expected estimation error measures the adversary's correctness by computing the expected distance between the true outcome x_c and the estimate x using a distance metric $d()$, for example the Euclidean distance or a metric that yields either 0 or 1 (in this case, the metric reduces to the adversary's probability of error). The expectation is computed over the posterior probability of the adversary's estimates x based on his observations o [Shokri et al. 2011].

$$priv_{\text{AEE}} \equiv \sum_x \widehat{Pr}(x|o) d(x, x_c)$$

5.6.2 Expectation of Distance Error

Similar to the adversary's expected estimation error, the expectation of distance error measures the expected distance error of an adversary, but over multiple timesteps K and location assignment hypotheses I [Hoh and Gruteser 2005]. It uses the probability of a hypothesis $p_i(k)$ and the total distance $d_i(k)$ between the correct user location assignment and hypothesis i for all users U at timestep k.

$$priv_{\text{EDE}} \equiv \frac{1}{|U|K} \sum_{k=1}^{K} \sum_{i=1}^{I} p_i(k) d_i(k)$$

5.6.3 Mean Squared Error

In statistical parameter estimations, a common goal is to minimize the mean squared error. As a privacy metric, the mean squared error describes the error between observations by the adversary o and the true outcome x_c, for example the error in the assignment of communication relationships [Oya et al. 2014], or the error in reconstructing user data in participatory sensing [Ganti et al. 2008].

$$priv_{\text{MSE}} \equiv \frac{1}{x_c} \sum_{x_c} \|x_c - o\|^2$$

5.6.4 Percentage Incorrectly Classified

This metric measures the percentage of incorrectly classified users or events I within the set of all users or events U, for example users that were incorrectly de-anonymized by the adversary [Narayanan and Shmatikov 2009], or events that were incorrectly classified in a smart metering scenario [Lisovich et al. 2010].

$$priv_{\text{PIC}} \equiv \frac{I}{U}$$

5.6.5 Health Privacy

Health privacy is a metric from genome privacy that captures privacy with regard to a specific disease [Humbert et al. 2013]. The metric assumes that S genetic variations contribute to the disease risk, where each variation contributes to a varying extent c_s. It is computed as the weighted, normalized sum over a base metric G_s^i which measures the privacy of each genetic variation. Base metrics can be normalized entropy (Section 5.1.4), normalized mutual information (Section 5.2.3), or expected estimation error (Section 5.6.1) [Humbert et al. 2013]. Depending on the base metric, health privacy measures a different kind of output; in the case of expected estimation error, health privacy measures the adversary's weighted average error.

$$priv_{\text{HLP}} \equiv \frac{1}{\sum_{s \in S} c_s} \sum_{s \in S} c_s G_s^i$$

5.7 Time

Time-based metrics focus on time as a resource that the adversary needs to spend to compromise users' privacy. Some time-based metrics measure the time until the adversary succeeds, assuming PETs will fail eventually, while others measure the time until the adversary's confusion, assuming PETs will succeed eventually.

5.7.1 Time until Adversary's Success

The most general time-based metric measures the time until the adversary's success [Wright et al. 2002]. It assumes that the adversary will succeed eventually, and is therefore an example of a pessimistic metric. This metric relies on a definition of success, and varies depending on how success is defined in a scenario.

For example, success in a communication system can be if the adversary identifies n out of N of the target's possible communication partners [Agrawal and Kesdogan 2003]. In an anonymization system that shuffles b messages in every round (a so-called batch mix), in which a single sender communicates with m recipients, and the security parameter is l, the metric computes the expected number of rounds until the adversary succeeds.

$$priv_{\text{TAI}} \equiv [m \cdot l(\sqrt{\frac{N-1}{N}(b-1)} + \sqrt{\frac{N-1}{N^2}(b-1) + \frac{m-1}{m}})]^2$$

Success can also be when the adversary first compromises a communication path [Johnson et al. 2013; Vratonjic et al. 2013]. In an onion routing system such as Tor [Dingledine et al. 2004], path compromise happens when the adversary controls all relays on a user's onion routing path.

5.7.2 Maximum Tracking Time

In location privacy, the adversary often aims to not only break privacy at a single point in time, but to track a target's location over time. The adversary's tracking ability is measured by the maximum tracking time, defined as the cumulative time that the size of the target t's anonymity set remains 1 [Sampigethaya et al. 2005].

$$priv_{\text{MTT}} \equiv \text{Cumulative time when } |AS_t| = 1$$

This metric tends to overestimate a target's privacy because it assumes that the adversary has to be completely certain, i.e., the anonymity set has to be of size 1, to be successful. In reality, however, an adversary may be capable to continue tracking despite a small number of users in the target's anonymity set.

In a smart metering scenario, the maximum tracking time describes the percentage of a time interval that the adversary can classify correctly [Lisovich et al. 2010].

5.7.3 Mean Time to Confusion

To avoid the maximum tracking time's overestimation of privacy, the mean time to confusion measures the time during which the adversary's uncertainty (measured using the entropy $H(X)$, Section 5.1.2) stays below a confusion threshold δ [Hoh et al. 2007].

$$priv_{\text{MTC}} \equiv \text{Time during which } H(X) < \delta$$

Instead of time to confusion, the metric can also measure the distance to confusion, i.e., the travel distance until the adversary's tracking uncertainty rises above the threshold.

5.8 Accuracy / Precision

Accuracy metrics quantify the accuracy of the adversary's estimate. Although it can be argued that the accuracy of an estimate is not correlated with privacy because it says nothing about the adversary's correctness or certainty [Shokri et al. 2011], inaccurate estimates can lead to higher privacy and are thus an important aspect of privacy.

5.8.1 Confidence Interval Width

According to the confidence interval width, the amount of privacy at $c\%$ confidence is given by the width of the interval for the adversary's estimate $[x_2, x_1]$ in which the true outcome x lies [Agrawal and Srikant 2000].

$$priv_{\text{CIW}} \equiv |x_2 - x_1| \text{ where } Pr(x_1 \leq x_{Adv} < x_2) = c/100$$

However, when publishing perturbed data, knowledge of the confidence interval width may allow reconstruction of the original distribution [Agrawal and Aggarwal 2001].

5.8.2 (t,p) Privacy Violation

In data mining, (t,p) privacy violation gives information whether the release of data can be seen as a privacy threat. Privacy is violated when an adversary can infer sensitive information S from public data D by building a classifier C with Bayes error ρ. The classifier is based on available training samples t, and its prediction accuracy is bounded by a privacy parameter p [Kantarcioğlu et al. 2004].

$$priv_{\text{TPP}} \equiv \rho(t;C) \leq \rho(t) - p, \text{ with } \rho(t) = \rho_{\{t;D,S\}} \text{ and } \rho(t;C) = \rho_{\{t;D,C,S\}}$$

5.8.3 Statistically Strong Event Unobservability

In wireless sensor networks, a privacy goal is to hide where in the network an event has occurred. Statistically strong event unobservability compares the message patterns in all parts of the network so that event locations are not revealed by a sudden burst of messages. Specifically, the metric requires that the distance between distributions (e.g., in terms of the difference between the areas under the cumulative distribution functions, $d(F_1, F_2)$) is smaller than α, and that the difference between the distribution parameters p is smaller than ϵ [Shao et al. 2008]. However, it only works when the distributions have a single parameter.

$$priv_{\text{SEU}} \equiv (\alpha, \epsilon), \text{ where } d(F_1, F_2) \leq \alpha \wedge (1-\epsilon)p_{F_1} \leq p_{F_2} \leq (1+\epsilon)p_{F_1}$$

5.8.4 Size of Uncertainty Region

In location privacy, the size of the uncertainty region denotes the minimal size of the region A to which an adversary can narrow down the position of a target user [Cheng et al. 2006].

$$priv_{\text{SUR}} \equiv Area(A)$$

5.8.5 Accuracy of Obfuscated Region

The accuracy of an obfuscated region indicates the required accuracy that a PET has to achieve to fulfill the user's requirements. This accuracy is computed based on the optimal accuracy provided by the used sensing technology r_{opt} and a user-specified minimum r_{min} [Ardagna et al. 2007].

$$priv_{\text{AOR}} \equiv \frac{r_{\text{opt}}^2}{r_{\text{min}}^2}$$

5.8.6 Coverage of Sensitive Region

The coverage of the sensitive region evaluates how a user's sensitive regions R_S overlap with the adversary's uncertainty region R_U [Cheng et al. 2006]. The metric is normalized to the area of the uncertainty region, so that it becomes 1 when R_U equals or is fully contained in R_S.

$$priv_{\text{CSR}} \equiv \frac{Area(R_S \cap R_U)}{Area(R_U)}$$

6 How to Select Privacy Metrics

Given the number and diversity of privacy metrics, selecting metrics for a given scenario can be difficult. We suggest a series of eight questions to guide the selection process. Answering each of the questions makes sure that all aspects of metric selection are considered. Where possible and appropriate, we point to metrics or groups of metrics that we associate with particular answers.

The first two questions ask about which aspects of privacy should be quantified (question 6.1), and which adversary types we need to protect against (question 6.2). Next, we suggest to consider which data sources need to be protected (question 6.3), and which input data are available to compute the metrics (question 6.4). We then move on to consider the requirements of the target audience (question 6.5) and which metrics have been used in related work (question 6.6). Finally, we suggest to check whether any of the selected metrics have flaws (question 6.7), and whether validated implementations for the metrics are available (question 6.8).

6.1 Suitable Output Measures?

Which aspects of privacy do we want to quantify? Do we want to give privacy guarantees, or is some loss of privacy acceptable?

The pool of potential metrics can be narrowed down by deciding which outputs we want to measure. In Section 4.4, we classify the output measures of privacy metrics into eight categories. Figure 1 and the *Output* column in Tables 6.8 and 2 list the output measure for each metric.

If the application scenario requires privacy guarantees in the sense that privacy properties can be proven to hold, the only viable choices for metrics are in the indistinguishability category. If the application instead calls for a quantification of privacy levels, metrics from the other categories are more suitable.

Instead of fixing a single output measure for a scenario, we recommend to measure several different outputs. Because none of the metrics measures 'privacy' directly, but only quantities assumed to be related to privacy, each additional output category gives information about an additional aspect of privacy.

For example, a study about location privacy by Shokri et al. [2011] used metrics from three different categories to measure the adversary's accuracy (confidence interval width, Section 5.8.1), uncertainty (entropy, Section 5.1.2), and error (expected estimation error, Section 5.6.1). Following our recommendation, this selection could be extended with a success metric that quantifies how likely it is for the adversary to succeed, or with a time metric that measures the time until the adversary's success. We might also add a second uncertainty metric that indicates the size of the crowd into which an individual can blend.

Besides including metrics from different categories, we recommend to select metrics that reflect the average case, the distribution of privacy values, and the worst case.

6.2 Adversary Models?

What are the characteristics of the adversary we consider? How do we incorporate the adversary and their knowledge?

We observed that papers presenting attacks against privacy tend to use metrics based on time, error, or the adversary's success probability, whereas papers presenting new PETs tend towards

accuracy, similarity, and indistinguishability metrics. In both cases, this is a convenient choice: most metrics in the first group have a stronger focus on the adversary, while the metrics in the second group emphasize the efficacy of the presented PET. However, as we have argued before, the measurement of privacy benefits when more aspects of privacy are measured. We therefore believe that both the 'attack' and 'defense' perspective can benefit from selecting metrics from the other side.

We also observed that different privacy domains make different assumptions about the adversary. For example, time-based metrics in communication systems measure the time until the adversary's success, whereas time-based metrics in location privacy measures the time until the adversary's confusion. This is a fundamental difference, and it is not obvious which flavor of the assumption holds in other privacy domains.

Care must be taken when choosing metrics that do not consider an adversary model. For example, most similarity/diversity metrics such as k-anonymity (Section 5.3) compute the level of privacy depending only on properties of the data. However, if the adversary happens to have relevant prior knowledge, the privacy level indicated by k is no longer accurate.

We found few metrics that explicitly consider the resources an adversary has to expend in order to succeed. Aside from time-based metrics, the only other metric considering resources is probability of path compromise (Section 5.5.1), where bandwidth and the number of nodes are the constrained resources. Resource-based metrics are an interesting area for future research, which means that if we consider a resource-constrained adversary, we will have to create new metrics.

6.3 Data Source?

Which data sources do we aim to protect?

We introduced four data sources in Section 4.2 – published, observable, re-purposed, or all other data. Depending on which data source needs protecting, different metrics apply. We summarize the primary data sources for each of the metrics in the *Primary data source* column in Tables 6.8 and 2.

Although in many scenarios one data source will be the main cause of concern, considering all four data sources reduces the likelihood that unforeseen events compromise the entire system. It also enables informed decisions about which privacy risks should be mitigated or accepted. In addition, considering all four data sources can emphasize the need for data minimization, because data that is not there does not need protection.

6.4 Availability of Input Data?

Which types of input data do we want to consider, and which are available in our scenario?

Input data refers to the information that is needed to compute a metric, such as the adversary's estimate, resources, and prior knowledge, the true outcome, or parameter values. If a certain kind of input data is not available or applicable in a scenario, we can disregard all metrics that need this input type. Similarly, if we explicitly want to consider a certain input, we can disregard metrics that do not use this input type. We describe different kinds of input data in Section 4.3 and show the kinds of input data for each metric in the *Inputs* column of Tables 6.8 and 2.

6.5 Target Audience?

What is the intended audience for our study? What are their expectations regarding the presentation of results, and do they understand the interpretations of our metrics?

An important consideration for the selection of metrics is the intended audience, especially with regard to laypeople and researchers in other academic disciplines.

Whenever results need to be communicated to laypeople, it is important to select metrics that can be understood easily. This does not mean that the formal definition of the metric has to be simplistic; rather, it means that the metric should have an intuitive interpretation, even if it simplifies the underlying technical details. However, we are not aware of user studies that evaluate how easily different metrics are understood by laypeople, or which interpretations help understanding.

Whenever metrics are intended to be used by researchers in other academic disciplines, it may be beneficial to use methods and terminology common in the respective discipline. Consider genome privacy as an example: in many areas of biology it is common to conduct statistical analyses; for non-privacy researchers in this field, metrics based on accuracy, error, or success will therefore be easier to understand and adopt than, say, metrics based on indistinguishability.

6.6 Related Work?

Which metrics are used by work that is related to ours, and would those metrics be suitable in our work as well? Which mathematical concepts or formalisms are used by others in our field? Which of these are already available in the tools we use?

To enable comparisons between different studies in the same privacy domain, it is useful to select those metrics that have already been used by related work, even if those metrics would otherwise not be the first choice. In addition, well-known metrics are likely to be more easily understood by other researchers in the same field.

A related consideration is expertise. Some metrics are conceptually difficult, and hard to use correctly. To reduce the risk of invalidating the results of an entire study, we recommend to select both comparatively simple metrics and more difficult ones.

6.7 Quality of Metrics?

Does any of the candidate metrics have known flaws? Is it feasible to conduct a study that verifies that candidate metrics indeed behave as we intend?

Even though it is desirable to work with high-quality metrics, few studies systematically evaluate the quality of privacy metrics. This means that information about metric quality is not readily available at the time of this writing. Even so, some metrics do have known weaknesses (which we have pointed out throughout Section 5) and should only be used with caution. If selecting known weak metrics, we recommend to use them in combination with other metrics to help offset the weaknesses.

If results about metric quality are not available for a particular privacy domain, it may be possible to conduct a small study to evaluate how candidate metrics perform.

6.8 Metric Implementations?

Are there implementations of the candidate metrics that we can use, or compare our implementation with?

Even when metrics are easy to understand, implementing them in a particular scenario can be difficult, and challenges can arise with unexpected aspects of a metric. For example, when implementing the entropy of an anonymity set, the challenge may not be entropy itself, but the propagation of anonymity set probabilities over multiple timesteps. Common challenges like this are likely to be solved to different degrees in different implementations. The more research groups use and validate an implementation, the higher the chance of detecting implementation errors. We therefore recommend to consider selecting metrics for which a validated implementation exists. Ultimately, only implementations that have been thoroughly validated can lead to consistent results across studies.

7 Future Research Directions

Despite the substantial body of research into privacy metrics presented in the previous section, there are a number of questions that merit further research.

7.1 Interdependent Privacy

Interdependent privacy refers to scenarios in which actions of one user affect the privacy of other users, for example in social networks [Thomas et al. 2010], location privacy [Vratonjic et al. 2013], or genome privacy [Humbert et al. 2013]. There are two options for measuring interdependent privacy. The first option is to measure how the value of an existing privacy metric changes when the degree of interdependency increases. The effect of interdependency can then be shown by comparing absolute values [Bloessl et al. 2015], or by computing a difference [Olteanu et al. 2014].

The second option is to create new metrics that explicitly consider interdependency. In this case, it can be beneficial to make use of metrics that measure the consequences one user's actions have on the privacy of another user. For example, this is done in game theory, where the widely used Helly metric [Vorob'ev 1977] assesses players' strategies in terms of their consequences which is the payoff for each player. We believe further research is needed to investigate the capabilities of these two options.

7.2 Privacy Attitudes and Behaviors

In this survey, we focused on technical privacy metrics and did not consider metrics that measure users' privacy attitudes, behaviors, or perception [Preibusch 2013]. User-assigned privacy or privacy

Output	Metric	Value range	high (H) or low (L) values indicate high privacy	Primary data source	Inputs				
					Adv. estimate	Adv. resources	True outcome	Prior knowledge	Parameters
Uncertainty	Anonymity Set Size	$[0, \infty]$	H	obs	x				
	Asymmetric Entropy	$[0, 1]$	H	obs, pub	x			x	
	Conditional entropy	$[0, \infty]$	H	obs, pub	x			x	
	Conditional privacy	$[1, \infty]$	H	obs, pub	x			x	
	Cross-entropy	$[0, \infty]$	H	pub	x		x		
	Cumulative entropy	$[0, \infty]$	H	obs	x				
	Degree of unlinkability	$[0, \infty]$	H	obs, pub	x			(x)	
	Entropy	$[0, H_0(X)]$	H	obs, pub	x				
	Entropy + Bayes	$[0, \infty]$	H	obs	x			x	
	Genomic Privacy	$[0, \infty]$	H	pub	x				x
	Inherent privacy	$[1, \lvert X \rvert]$	H	obs, pub	x				
	Max-entropy (Hartley)	$[0, \infty]$	H	obs, pub	x				
	Min-entropy	$[0, \infty]$	L	obs, pub	x				
	Normalized conditional entropy	$[0, 1]$	H	obs, pub	x			x	
	Normalized entropy	$[0, 1]$	H	obs, pub	x				
	Protection Level	$[0, \infty]$	H	obs	x				x
	Quantiles on entropy	$[0, H_0(X)]$	H	obs, pub	x				x
	Rényi entropy	$[0, \infty]$	H	obs, pub	x				x
	User-centric privacy	$[0, H_0(U)]$	H	obs	x				x
Information Gain	Amount of leaked information	$[0, \infty]$	L	pub, oth			x		
	Conditional Mutual Information	$[0, \infty]$	L	obs, pub	x		x	x	
	Conditional privacy loss	$[0, 1]$	L	obs, pub	x		x		
	Increase in adversary's belief	true, false	L	obs, pub	x			x	x
	Information Surprisal	$]0, \infty]$	L	pub	x		x		
	Maximum information leakage	$[0, \infty]$	L	obs, pub	x				
	Mutual information	$[0, \infty]$	L	obs, pub	x		x		
	Normalized mutual information	$[0, 1]$	H	obs, pub	x		x		
	Pearson's correlation coefficient	$[0, 1]$	L	obs, rep			x		
	Privacy Score	$[0, \infty]$	L	pub					x
	Reduction in observable features	$[0, 1]$	L	obs, rep			x		
	Relative entropy	$[0, \infty]$	H	obs, pub	x		x		
	(Relative) Loss of anonymity	$[0, H(X)]$	L	obs	x		x	(x)	
	System anonymity level	$[0, \infty]$	H	obs	x		x		
Similarity	(α,k)-anonymity	k: $[0, \infty]$, α: $[0, 1]$	k: H, α: L	pub					x
	(c,t)-isolation	$[0, \infty]$	H	pub	x		x		x
	Cluster similarity	$[0, 1]$	L	obs, rep			x		
	Coefficient of determination R^2	$[0, 1]$	L	obs, rep			x		
	(ϵ,m)-anonymity	$\epsilon : [0, 1]$, m: $[1, \infty]$	ϵ: H, m: H	pub					x
	Haplotype-SNP-test	true, false	H	pub					x
	Historical k-Anonymity	$[0, \infty]$	H	obs			x		x
	k-anonymity	$[1, \lvert D \rvert]$	H	pub					x
	(k,e)-anonymity	$[0, \infty]$	H	pub					x
	ℓ-diversity	$[0, \infty]$	H	pub					x
	m-invariance	$[0, \infty]$	H	pub					x
	Multirelational k-anonymity	$[0, \infty]$	H	pub			x		x
	t-closeness	$[0, \infty]$	L	pub			x		x
	Normalized variance	$[0, 1]$	H	pub			x		
	(X,Y)-privacy	$]0, 1]$	L	pub			x		x

Table 1: Privacy Metrics (1): Uncertainty, Information Gain/Loss, and Similarity/Diversity Outputs

Output	Metric	Value range	high (H) or low (L) values indicate high privacy	Primary data source	Inputs: Adv. estimate	Adv. resources	True outcome	Prior knowledge	Parameters
Time	Maximum tracking time	$[0, \infty]$	L	obs	x				
	Mean time to confusion	$[0, \infty]$	L	obs	x				x
	Time until adversary's success	$[0, \infty]$	H	obs	x		x		(x)
Indistinguishability	Approximate differential privacy	ϵ: $[0, \infty]$, δ: $[0, \infty]$	ϵ: L, δ: L	pub			x		x
	Computational differential privacy	$[0, \infty]$	L	pub	x	x	x		x
	Cryptographic game	true, false	H	obs	x		x		x
	Differential privacy	$[0, \infty]$	L	pub			x		x
	Distributed differential privacy	ϵ: $[0, \infty]$, δ: $[0, \infty]$	ϵ: L, δ: L	pub, rep			x		x
	Distributional privacy	$[0, \infty]$	L	pub, rep			x		x
	Geo-indistinguishability	$[0, \infty]$	L	obs			x		x
	Information privacy	true, false	H	obs	x				x
	Observational equivalence	true, false	H	obs	x		x		
	Unconditional / computational privacy	true, false	L	obs	x		x		x
Success	Adversary's success rate	$[0, 1]$	L	obs	x		x		(x)
	(d,γ)-privacy	$[0, 1]$	L	obs	x			x	x
	Degrees of Anonymity	$[0, 1]$	L	obs	x		x		x
	δ-presence	$[0, 1]$	L	pub	x			x	x
	Hiding property	$[0, 1]$	L	obs	x				x
	Privacy breach level	$[0, 1]$	L	obs	x			x	x
	Probability of path compromise	$[0, 1]$	L	obs	x	x	x		
	Adversary's expected estimation error	$[0, 1]$	L	obs	x		x		
Error	Expectation of distance error	$[0, \infty]$	H	obs	x		x		
	Mean Squared Error	$[0, \infty]$	H	obs	x		x		
	Percentage incorrectly classified	$[0, 1]$	H	obs, rep	x		x		
Accuracy	Accuracy of obfuscated region	$[0, 1]$	L	obs					x
	Confidence interval width	$[0, \infty]$	H	pub, obs	x				x
	Coverage of sensitive region	$[0, 1]$	L	obs	x				x
	Size of uncertainty region	$[0, \infty]$	H	obs	x				
	Statistically strong event unobservability	$[0, \infty]$	L	obs	x				x
	(t,p) privacy violation	$[0, 1]$	L	pub	x		x	x	x

Table 2: Privacy Metrics (2): Time, Indistinguishability, Adversary's Success Probability, Error, and Accuracy/Precision Outputs

risk scores vary greatly in how information is collected from the user. For example, some studies measure users' perception of privacy risks or privacy attitudes on Likert scales [Acquisti et al. 2003; Achara et al. 2014]. Others require users to label sensitive data [Zhang et al. 2011], assign privacy scores to their credentials [Yao et al. 2008], or configure existing mechanisms according to their privacy needs [Xiao and Tao 2006]. Some studies work with risk attitudes that are inferred from user actions via machine learning [Akcora et al. 2012].

Some metrics in our survey combine a technical metric with parameters that are specified by users to reflect their preferences, for example user-centric privacy (Section 5.1.13), coverage of sensitive region (Section 5.8.6), or privacy score (Section 5.2.12). In general, however, it is an open question how best to integrate user attitudes, behaviors, or perceptions with technical metrics. In addition, it is unclear whether this integration is generally useful, and which scenarios would benefit most.

7.3 Aggregating Metrics

In scenarios with a large number of entities, such as thousands of genomic variations or users in a communication system, it can be beneficial to aggregate metrics. Some metrics in our survey attempt to do this, for example cumulative entropy (Section 5.1.10), genomic privacy (Section 5.1.11), health privacy (Section 5.6.5), or expected estimation error (Section 5.6.1). All of these metrics are based on an addition of privacy values. Their results are a sum (cumulative entropy, genomic privacy), a weighted arithmetic mean (health privacy), or an expected value (expected estimation error). However, depending on the distribution of the underlying population, the arithmetic mean may lead to biased results [Mashey 2004]. In some situations, a geometric mean is preferable because it assumes a log-normal, rather than normal, distribution, and is less biased by outlier values [Citron et al. 2006]. However, in the field of privacy measurement it is not clear what these situations are. We therefore believe that privacy research would benefit from a rigorous study of ways to aggregate metrics.

Another option to aggregate privacy values is visualization. When metrics are visualized, a common option is to display averages – the same strategy as with aggregate metrics. However, more sophisticated plot types can highlight issues such as fairness that are hidden when averages are used. For example, box plots display the smallest and largest privacy values as well as the first, second, and third quartile; and violin plots add kernel density plots to visualize the distribution of privacy values. These plots give more information than aggregate metrics; however, it is unclear how aggregate metrics can be designed so that the benefits of these plots are preserved.

7.4 Combining Metrics

Whereas the aggregation of metrics considers values of the same privacy metric for many entities, the combination of metrics considers values of different privacy metrics for one entity. Combining different metrics can be useful if the combination retains the strengths of each metric while reducing their weaknesses. It can also simplify interpretation to express the performance of a PET with a single number. Metrics in our survey use three methods to combine metrics: adding sensitivity scores, normalizing metrics, and extending metrics to new contexts.

Metrics that combine a sensitivity score with a technical metric are user-centric privacy (using a linear combination, Section 5.1.13) and privacy score (using sensitivity as a weighting factor, Section 5.2.12). As mentioned in Section 7.2 above, it is not clear how sensitivity scores and technical metrics can best be combined. In addition, it is not clear whether the resulting values have a meaningful interpretation.

Metrics that combine two technical metrics typically use one metric to normalize another, for example normalized entropy (Section 5.1.4), normalized mutual information (Section 5.2.3), or reduction in observable features (Section 5.2.11). Normalization can make it easier to interpret privacy measurements, but for some metrics, is is not clear if and how they can be normalized, or which normalization method works best.

Metrics that adapt a privacy metric so that it can be used in a new context are computational differential privacy which adapts differential privacy (Section 5.4.7) to a new adversary type, and entropy combined with Bayesian belief tables to apply entropy across multiple time-steps (Section 5.1.2). These innovative metrics raise two questions: first, whether their mechanisms can extend the range of use for other metrics as well, and second, whether there are other mechanisms that can be used in a similar way to adapt existing metrics to new use cases.

7.5 Quality of Metrics

We presented a number of quality indicators for privacy metrics in Section 2. While there is a general consensus that high-quality metrics should be used, there is no consensus what exactly constitutes high quality and how it should be measured. As a result, there are few studies investigating the quality of privacy metrics. For example, in a previous study, we systematically compared 22 metrics for genome privacy and found that metrics varied greatly with regard to consistency and monotonicity [Wagner 2015]. Although our study yielded good results for a selection of privacy metrics in one specific scenario, it was limited in terms of the scenario, quality indicators, and number of privacy metrics. It is unclear whether the results of our study would hold in general, and therefore we believe that more studies are needed that rigorously evaluate the quality of privacy metrics.

8 Conclusion

In this survey we presented a comprehensive review of privacy metrics. We described and discussed a selection of over eighty privacy metrics using examples from six different privacy domains.

To structure the complex landscape of privacy metrics, we introduced a categorization based on the aspect of privacy they measure, their required inputs, and the type of data that needs protection. In addition, we highlighted topics where we believe additional work on privacy metrics is needed. This includes research toward the combination and aggregation of privacy metrics as well as the field of interdependent privacy.

Finally, we presented a method on how to choose privacy metrics based on eight questions that help identify the right privacy metrics for a given scenario. Most importantly, we argue for the selection of multiple metrics to cover multiple aspects of privacy. We believe that this systematization will serve as a reference guide for privacy metrics that allows informed choices of privacy metrics and thus serves as a useful toolbox for privacy researchers.

References

Jagdish Prasad Achara, Mathieu Cunche, Vincent Roca, and Aurélien Francillon. WifiLeaks: Underestimated Privacy Implications of the ACCESS_WIFI_STATE Android Permission. In *Proc. 7th ACM Conf. on Security and Privacy in Wireless and Mobile Networks (WiSec 2014)*, pages 231–236, Oxford, UK, July 2014. ACM. ISBN 978-1-4503-2972-9.

Alessandro Acquisti, Roger Dingledine, and Paul Syverson. On the Economics of Anonymity. In *Proc. 7th Int. Financial Cryptography Conf (FC03)*, pages 84–102, Gosier, Guadeloupe, January 2003. Springer.

Charu C. Aggarwal. On k-Anonymity and the Curse of Dimensionality. In *Proc. 31st Int. Conf. on Very Large Data Bases (VLDB 2005)*, pages 901–909, Trondheim, Norway, September 2005. VLDB Endowment.

Dakshi Agrawal and Charu C. Aggarwal. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In *Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems (PODS 2001)*, pages 247–255, Santa Barbara, CA, USA, 2001 2001. ACM.

Dakshi Agrawal and Dogan Kesdogan. Measuring Anonymity: The Disclosure Attack. *IEEE Security & Privacy*, 1(6):27–34, November 2003. ISSN 1540-7993.

Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving Data Mining. In *Proc. 2000 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'00)*, pages 439–450, Dallas, TX, USA, May 2000. ACM. ISBN 1-58113-217-4.

Cuneyt Gurcan Akcora, Barbara Carminati, and Elena Ferrari. Privacy in Social Networks: How Risky is Your Social Graph? In *Proc IEEE 28th Int. Conf. on Data Engineering (ICDE'12)*, pages 9–19, Washington, DC, USA, April 2012. IEEE.

James Alexander and Jonathan Smith. Engineering Privacy in Public: Confounding Face Recognition. In *Proc. 3rd Int. Workshop on Privacy Enhancing Technologies (PET 2003)*, LNCS 2760, pages 88–106, Dresden, Germany, March 2003. Springer. ISBN 978-3-540-20610-1, 978-3-540-40956-4.

Christer Andersson and Reine Lundin. On the Fundamentals of Anonymity Metrics. In *Proc. 3rd IFIP Int. Summer School on The Future of Identity in the Information Society*, pages 325–341, Karlstad, Sweden, August 2008. Springer.

Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-Indistinguishability: Differential Privacy for Location-Based Systems. In *Proc. 20th ACM Conf. on Computer and Communications Security (CCS'13)*, pages 901–914, Berlin, Germany, November 2013. ACM. ISBN 978-1-4503-2477-9.

Myrto Arapinis, Loretta Mancini, Eike Ritter, Mark Ryan, Nico Golde, Kevin Redon, and Ravishankar Borgaonkar. New Privacy Issues in Mobile Telephony: Fix and Verification. In *Proc. 19th*

ACM Conf. on Computer and Communications Security (CCS'12), pages 205–216, Raleigh, NC, USA, October 2012. ACM. ISBN 978-1-4503-1651-4.

Claudio Agostino Ardagna, Marco Cremonini, Ernesto Damiani, S. De Capitani Di Vimercati, and Pierangela Samarati. Location Privacy Protection Through Obfuscation-based Techniques. In *Data and Applications Security XXI: 21st Annu. IFIP Working Conf. on Data and Applications Security*, pages 47–60, Redondo Beach, CA, USA, July 2007. Springer.

Erman Ayday, Jean Louis Raisaro, and Jean-Pierre Hubaux. Personal Use of the Genomic Data: Privacy vs. Storage Cost. In *Proc. IEEE Global Communications Conf. (GLOBECOM 2013)*, pages 2723–2729, Atlanta, GA, USA, December 2013a. IEEE.

Erman Ayday, Jean Louis Raisaro, Jean-Pierre Hubaux, and Jacques Rougemont. Protecting and Evaluating Genomic Privacy in Medical Tests and Personalized Medicine. In *Proc. 12th ACM Workshop on Workshop on Privacy in the Electronic Society (WPES'13)*, pages 95–106, Berlin, Germany, November 2013b. ACM. ISBN 978-1-4503-2485-4.

Michael Backes, Stefan Lorenz, Matteo Maffei, and Kim Pecina. Anonymous Webs of Trust. In *Proc. 10th Int. Symp. on Privacy Enhancing Technologies (PETS 2010)*, LNCS 6205, pages 130–148, Berlin, Germany, January 2010. Springer. ISBN 978-3-642-14526-1, 978-3-642-14527-8.

Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. Wherefore Art Thou R3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography. In *Proc. 16th Int. Conf. on World Wide Web (WWW'07)*, pages 181–190, Banff, Canada, May 2007. ACM.

Elisa Bertino, Dan Lin, and Wei Jiang. A Survey of Quantification of Privacy Preserving Data Mining Algorithms. In *Privacy-Preserving Data Mining: Models and Algorithms*, number 34 in Advances in Database Systems, chapter 8, pages 183–205. Springer, July 2008. ISBN 978-0-387-70991-8, 978-0-387-70992-5.

Claudio Bettini, X. Sean Wang, and Sushil Jajodia. Protecting Privacy Against Location-based Personal Identification. In *2nd VLDB Workshop on Secure Data Management (SDM 2005)*, LNCS 3674, pages 185–199, Trondheim, August, September 2005. Springer.

Bastian Bloessl, Christoph Sommer, Falko Dressler, and David Eckhoff. The scrambler attack: A robust physical layer attack on location privacy in vehicular networks. In *2015 International Conference on Computing, Networking and Communications (ICNC)*, pages 395–400, February 2015.

Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Prakash Panangaden. Anonymity Protocols as Noisy Channels. In *Proc. 3rd Int. Symp. Trustworthy Global Computing (TGC'2007)*, pages 281–300, Sophia-Antipolis, France, November 2007. Springer.

David Chaum. The Dining Cryptographers Problem: Unconditional Sender and Recipient Untraceability. *Journal of Cryptology*, 1(1):65–75, January 1988. ISSN 0933-2790, 1432-1378.

Shuchi Chawla, Cynthia Dwork, Frank McSherry, Adam Smith, and Hoeteck Wee. Toward Privacy in Public Databases. In *Proc. 2nd Int. Conf. on Theory of Cryptography (TCC'05)*, pages 363–385, Cambridge, MA, USA, February 2005. Springer.

Terence Chen, Abdelberi Chaabane, Pierre Ugo Tournoux, Mohamed-Ali Kaafar, and Roksana Boreli. How Much Is Too Much? Leveraging Ads Audience Estimation to Evaluate Public Profile Uniqueness. In *Proc. 13th Int. Symp. on Privacy Enhancing Technologies (PETS 2013)*, LNCS 7981, pages 225–244, Bloomington, IN, USA, July 2013. Springer. ISBN 978-3-642-39076-0, 978-3-642-39077-7.

Xihui Chen and Jun Pang. Measuring Query Privacy in Location-based Services. In *Proc. 2nd ACM Conf. on Data and Application Security and Privacy (CODASPY'12)*, pages 49–60, San Antonio, TX, USA, February 2012. ACM. ISBN 978-1-4503-1091-8.

Reynold Cheng, Yu Zhang, Elisa Bertino, and Sunil Prabhakar. Preserving User Location Privacy in Mobile Data Management Infrastructures. In *Proc. 6th Int. Workshop on Privacy Enhancing Technologies (PET 2006)*, LNCS 4258, pages 393–412, Cambridge, UK, June 2006. Springer. ISBN 978-3-540-68790-0, 978-3-540-68793-1.

Daniel Citron, Adham Hurani, and Alaa Gnadrey. The Harmonic or Geometric Mean: Does It Really Matter? *SIGARCH Comput. Archit. News*, 34(4):18–25, September 2006. ISSN 0163-5964.

Sebastian Clauß and Stefan Schiffner. Structuring Anonymity Metrics. In *Proc. 13th ACM Conf. on Computer and Communications Security 2006 (CCS'06): 2nd ACM Workshop on Digital Identity Management (DIM'06)*, pages 55–62, Alexandria, VA, USA, October 2006. ACM. ISBN 1-59593-547-9.

Aaron R. Coble. Formalized Information-Theoretic Proofs of Privacy Using the HOL4 Theorem-Prover. In *Proc 8th Int. Symp. on Privacy Enhancing Technologies (PETS 2008)*, pages 77–98, Leuven, Belgium, July 2008. Springer.

Council of Europe. *European Convention on Human Rights*. Council of Europe, Strasbourg, July 2010. ISBN 9789287169297.

Stéphanie Delaune, Steve Kremer, and Mark Ryan. Verifying Privacy-type properties of electronic voting protocols. *Journal of Computer Security*, 17(4):435–487, December 2009.

Yuxin Deng, Jun Pang, and Peng Wu. Measuring Anonymity with Relative Entropy. In *Proc. 8th Int. Workshop on Formal Aspects in Security and Trust (FAST 2011)*, pages 65–79, Leuven, Belgium, September 2007. Springer.

Claudia Diaz, Stefaan Seys, Joris Claessens, and Bart Preneel. Towards Measuring Anonymity. In *Proc. 3rd Int. Workshop on Privacy Enhancing Technologies (PET 2003)*, LNCS 2482, pages 54–68, Dresden, Germany, March 2003. Springer. ISBN 978-3-540-00565-0, 978-3-540-36467-2.

Claudia Diaz, Carmela Troncoso, and George Danezis. Does Additional Information Always Reduce Anonymity? In *Proc. 6th ACM Workshop on Privacy in Electronic Society (WPES '07)*, pages 72–75, Alexandria, VA, USA, October 2007. ACM. ISBN 978-1-59593-883-1.

Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor: The Second-Generation Onion Router. In *Proc. 13th USENIX Security Symp. (Security'04)*, pages 1–17, San Diego, CA, USA, August 2004. USENIX.

Flávio du Pin Calmon and Nadia Fawaz. Privacy Against Statistical Inference. In *Proc. 50th Annu. Allerton Conf. on Communication, Control, and Computing (Allerton 2012)*, pages 1401–1408, Monticello, IL, USA, October 2012. IEEE.

Matt Duckham and Lars Kulik. A formal model of obfuscation and negotiation for location privacy. In *Pervasive computing*, pages 152–170. Springer, 2005.

Cynthia Dwork. Differential Privacy. In *Proc. 33rd Int. Colloq. on Automata, Languages and Programming (ICALP 2006)*, LNCS 4052, pages 1–12, Venice, Italy, July 2006. Springer.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our Data, Ourselves: Privacy via Distributed Noise Generation. In *Proc. 25th Int. Cryptology Conf. (EUROCRYPT 2006)*, pages 486–503, St. Petersburg, Russia, June 2006. Springer.

Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil Vadhan. On the Complexity of Differentially Private Data Release: Efficient Algorithms and Hardness Results. In *Proc. 41st Annu. ACM Symp. on Theory of Computing (STOC'09)*, pages 381–390, Bethesda, MD, USA, May 2009. ACM.

European Parliament. Directive 95/46/EC. *Official Journal*, L 281:0031–0050, November 1995. URL `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML`.

Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke. Privacy Preserving Mining of Association Rules. *Information Systems*, 29(4):343–364, June 2004.

Matthias Franz, Bernd Meyer, and Andreas Pashalidis. Attacking Unlinkability: The Importance of Context. In *Proc. 7th Int. Symp. on Privacy Enhancing Technologies (PETS 2007)*, LNCS 4776, pages 1–16, Ottawa, Canada, June 2007. Springer. ISBN 978-3-540-75550-0, 978-3-540-75551-7.

Julien Freudiger, Maxim Raya, Márk Félegyházi, Panos Papadimitratos, and Jean-Pierre Hubaux. Mix-Zones for Location Privacy in Vehicular Networks. In *Proc. 1st Int. Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS 2007)*, Vancouver, Canada, August 2007. ICST.

Julien Freudiger, Mohammad Hossein Manshaei, Jean-Pierre Hubaux, and David C. Parkes. On Non-cooperative Location Privacy: A Game-theoretic Analysis. In *Proc. 16th ACM Conf. on Computer and Communications Security (CCS'09)*, pages 324–337, Chicago, IL, USA, November 2009. ACM. ISBN 978-1-60558-894-0.

Benjamin Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Computing Surveys (CSUR)*, 42(4):14, June 2010.

Raghu K. Ganti, Nam Pham, Yu-En Tsai, and Tarek F. Abdelzaher. PoolView: stream privacy for grassroots participatory sensing. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 281–294. ACM, 2008.

Benedikt Gierlichs, Carmela Troncoso, Claudia Diaz, Bart Preneel, and Ingrid Verbauwhede. Revisiting a Combinatorial Approach Toward Measuring Anonymity. In *Proc. 7th ACM Workshop on Privacy in the Electronic Society (WPES'08)*, pages 111–116, Alexandria, VA, USA, October 2008. ACM. ISBN 978-1-60558-289-4.

Philippe Golle and Kurt Partridge. On the anonymity of home/work location pairs. In *Pervasive Computing*, pages 390–397. Springer, 2009.

Angèle Hamel, Jean-Charles Grégoire, and Ian Goldberg. The Misentropists: New Approaches to Measures in Tor. Technical report, Technical Report 2011-18, Cheriton School of Computer Science, University of Waterloo, 2011. URL `http://cacr.uwaterloo.ca/techreports/2011/cacr2011-18.pdf`.

Jens Hermans, Andreas Pashalidis, Frederik Vercauteren, and Bart Preneel. A new RFID Privacy Model. In *Proc. 16th Symp. on Research in Computer Security (ESORICS 2011)*, LNCS 6879, pages 568–587, Leuven, Belgium, September 2011. Springer.

Thomas S. Heydt-Benjamin, Hee-Jin Chae, Benessa Defend, and Kevin Fu. Privacy for Public Transportation. In George Danezis and Philippe Golle, editors, *Privacy Enhancing Technologies*, number 4258 in Lecture Notes in Computer Science, pages 1–19. Springer Berlin Heidelberg, January 2006. ISBN 978-3-540-68790-0 978-3-540-68793-1.

Baik Hoh and Marco Gruteser. Protecting Location Privacy Through Path Confusion. In *Proc. 1st Int. Conf. on Security and Privacy for Emerging Areas in Communications Networks, (SecureComm 2005)*, pages 194–205, Athens, Greece, September 2005. IEEE.

Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. Preserving Privacy in GPS Traces via Uncertainty-Aware Path Cloaking. In *Proc. 14th ACM Conf. on Computer and Communications Security (CCS'07)*, pages 161–171, Alexandria, VA, USA, October 2007. ACM. ISBN 978-1-59593-703-2.

Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C. Pierce, and Aaron Roth. Differential Privacy: An Economic Method for Choosing Epsilon. *arXiv:1402.3329 [cs]*, February 2014.

Dominic Hughes and Vitaly Shmatikov. Information Hiding, Anonymity and Privacy: A Modular Approach. *ACM Journal of Computer Security*, 12(1):3–36, January 2004.

Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy. In *Proc. 20th ACM Conf. on Computer and Communications Security (CCS'13)*, pages 1141–1152, Berlin, Germany, November 2013. ACM. ISBN 978-1-4503-2477-9.

Márk Jelasity and Kenneth P. Birman. Distributional Differential Privacy for Large-scale Smart Metering. In *Proc. 2nd ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'14)*, pages 141–146, Salzburg, Austria, June 2014. ACM.

Aaron Johnson, Chris Wacek, Rob Jansen, Micah Sherr, and Paul Syverson. Users Get Routed: Traffic Correlation on Tor by Realistic Adversaries. In *Proc. 20th ACM Conf. on Computer and Communications Security (CCS'13)*, page 337–348, Berlin, Germany, November 2013. ACM. ISBN 978-1-4503-2477-9.

Ari Juels and Stephen A. Weis. Defining Strong Privacy for RFID. *ACM Trans. Inf. Syst. Secur.*, 13(1):7:1–7:23, November 2009. ISSN 1094-9224.

Georgios Kalogridis, Costas Efthymiou, Stojan Z. Denic, Tim A. Lewis, and Rafael Cepeda. Privacy for Smart Meters: Towards Undetectable Appliance Load Signatures. In *Proc. 1st Int. Conf. on Smart Grid Communications (SmartGridComm 2010)*, pages 232–237, Gaithersburg, MD, USA, October 2010. IEEE.

Murat Kantarcioğlu, Jiashun Jin, and Chris Clifton. When do Data Mining Results Violate Privacy? In *Proc. 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and data Mining (KDD'04)*, pages 599–604, Seatlle, WA, USA, August 2004. ACM.

Douglas J. Kelly, Richard A. Raines, Michael R. Grimaila, Rusty O. Baldwin, and Barry E. Mullins. A Survey of State-of-the-art in Anonymity Metrics. In *Proc. 15th ACM Conf. on Computer and Communications Security 2008 (CCS'08): 1st Workshop on Network Data Anonymization (NDA'08)*, pages 31–40, Alexandria, VA, USA, October 2008. ACM. ISBN 978-1-60558-301-3.

Dogan Kesdogan, Jan Egner, and Roland Büschkes. Stop- and- Go-MIXes Providing Probabilistic Anonymity in an Open System. In *Proc. 2nd Int. Workshop on Information Hiding (IH'98)*, LNCS 1525, pages 83–98, Portland, OR, USA, April 1998. Springer. ISBN 978-3-540-65386-8, 978-3-540-49380-8.

Daniel Kifer and Ashwin Machanavajjhala. No Free Lunch in Data Privacy. In *Proc. 2011 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'11)*, pages 193–204, Athens, Greece, June 2011. ACM.

Younghun Kim, E. C. Ngai, and Mani B. Srivastava. Cooperative State Estimation for Preserving Privacy of User Behaviors in Smart Grid. In *Proc. 2nd IEEE Int. Conf. on Smart Grid Communications (SmartGridComm 2011)*, pages 178–183, Brussels, Belgium, October 2011. IEEE.

John Krumm. A Survey of Computational Location Privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, August 2009.

Lifeng Lai, Siu-Wai Ho, and Vincent H. Poor. Privacy-Security Trade-Offs in Biometric Security Systems—Part I: Single Use Case. *IEEE Trans. on Information Forensics Security*, 6(1):122–139, March 2011. ISSN 1556-6013.

Jiexing Li, Yufei Tao, and Xiaokui Xiao. Preservation of Proximity Privacy in Publishing Numerical Sensitive Data. In *Proc. 2008 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2008)*, pages 473–486, Vancouver, Canada, June 2008. ACM.

Ninghui Li, Tiancheng Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *Proc. IEEE 23rd Int. Conf. on Data Engineering (ICDE 2007)*, pages 106–115, Istanbul, Turkey, April 2007. IEEE.

Zhen Lin, Michael Hewett, and Russ B. Altman. Using Binning to Maintain Confidentiality of Medical Data. In *Proc. AMIA Symp. (AMIA 2002)*, pages 454–458, San Antonio, TX, USA, November 2002.

Mikhail A. Lisovich, Deirdre K. Mulligan, and Stephen B. Wicker. Inferring personal information from demand-response systems. *Security & Privacy, IEEE*, 8(1):11–20, 2010.

Kun Liu and Evimaria Terzi. A Framework for Computing the Privacy Scores of Users in Online Social Networks. *ACM Transactions on Knowledge Discovery from Data*, 5(1):6:1–6:30, December 2010. ISSN 1556-4681.

Zhendong Ma, Frank Kargl, and Michael Weber. Measuring long-term location privacy in vehicular communication systems. *Elsevier Computer Communications*, 33(12):1414–1427, March 2010. ISSN 0140-3664.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3:1–3:52, March 2007. ISSN 1556-4681.

John R. Mashey. War of the Benchmark Means: Time for a Truce. *SIGARCH Comput. Archit. News*, 32(4):1–14, September 2004. ISSN 0163-5964.

Stephen McLaughlin, Patrick McDaniel, and William Aiello. Protecting Consumer Privacy from Electric Load Monitoring. In *Proc. 18th ACM Conf. on Computer and Communications Security (CCS'11)*, pages 87–98, Chicago, IL, USA, October 2011. ACM. ISBN 978-1-4503-0948-6.

Frank D. McSherry. Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis. In *Proc. 2009 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2004)*, pages 19–30, Providence, RI, USA, Junge 2009. ACM.

Srujana Merugu and Joydeep Ghosh. Privacy-preserving Distributed Clustering using Generative Models. In *Proc. 3rd Int. Conf. on Data Mining (ICDM'03)*, pages 211–218, Melbourne, FL, USA, November 2003. IEEE.

Ilya Mironov, Omkant Pandey, Omer Reingold, and Salil Vadhan. Computational Differential Privacy. In *Proc. 29th Annu. Int. Cryptology Conf. (CRYPTO 2009)*, LNCS 5677, pages 126–142, Santa Barbara, CA, USA, August 2009. Springer.

Steven J. Murdoch. Quantifying and Measuring Anonymity. In *Proc. 18th Symp. on Research in Computer Security (ESORICS 2013), 7th Int. Workshop on Autonomous and Spontaneous Security (SETOP 2013)*, LNCS, pages 3–13, Rhul, UK, September 2013. Springer. ISBN 978-3-642-54567-2, 978-3-642-54568-9.

Steven J. Murdoch and Robert N. M. Watson. Metrics for Security and Performance in Low-Latency Anonymity Systems. In *Proc. 8th Int. Symp. on Privacy Enhancing Technologies (PETS 2008)*, LNCS 5134, pages 115–132, Leuven, Belgium, July 2008. Springer. ISBN 978-3-540-70629-8, 978-3-540-70630-4.

Arvind Narayanan and Vitaly Shmatikov. Robust De-anonymization of Large Sparse Datasets. In *Proc. 2008 IEEE Symp. on Security and Privacy (S&P 2008)*, pages 111–125, May, Oakland, CA, USA 2008. IEEE.

Arvind Narayanan and Vitaly Shmatikov. De-anonymizing Social Networks. In *Proc. 2009 30th IEEE Symp. on Security and Privacy (S&P 2009)*, pages 173–187, Oakland, CA, USA, May 2009. IEEE.

Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. Hiding the presence of individuals from shared databases. In *Proc. 2007 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2007)*, pages 665–676, Beijing, China, June 2007. ACM.

Mehmet Ercan Nergiz, Chris Clifton, and Ahmet Erhan Nergiz. MultiRelational k-Anonymity. *IEEE Trans. on Knowledge Data Engineering*, 21(8):1104–1117, August 2009.

Helen Nissenbaum. Privacy as Contextual Integrity. *Washington Law Review*, 79(1):119–158, 2004.

Stanley R. M. Oliveira and Osmar R. Zaïane. Privacy Preserving Clustering By Data Transformation. In *Proc. 18th Brazilian Symp. on Databases (SBBD'2003)*, pages 304–318, Manaus, Brazil, October 2003.

Alexandra-Mihaela Olteanu, Kévin Huguenin, Reza Shokri, and Jean-Pierre Hubaux. Quantifying the Effect of Co-location Information on Location Privacy. In *Proc. 14th Int. Symp. on Privacy Enhancing Technologies (PETS 2014)*, LNCS 8555, pages 184–203, Amsterdam, Netherlands, July 2014. Springer. ISBN 978-3-319-08505-0, 978-3-319-08506-7.

Simon Oya, Carmela Troncoso, and Fernando Pérez-González. Do Dummies Pay Off? Limits of Dummy Traffic Protection in Anonymous Communications. In *Proc. 14th Int. Symp. on Privacy Enhancing Technologies (PETS 2014)*, LNCS 8555, pages 204–223, Amsterdam, Netherlands, July 2014. Springer. ISBN 978-3-319-08505-0, 978-3-319-08506-7.

Sören Preibusch. Guide to measuring privacy concern: Review of survey and observational instruments. *International Journal of Human-Computer Studies*, 71(12):1133–1143, December 2013. ISSN 1071-5819.

Vibhor Rastogi, Dan Suciu, and Sungho Hong. The Boundary Between Privacy and Utility in Data Publishing. In *Proc. 33rd Int. Conf. on Very Large Data Bases (VLDB 2007)*, pages 531–542, September, Vienna, Austria 2007. VLDB Endowment.

Michael K. Reiter and Aviel D. Rubin. Crowds: Anonymity for Web Transactions. *ACM Transactions on Information and System Security (TISSEC)*, 1(1):66–92, November 1998.

Pierangela Samarati and Latanya Sweeney. Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. In *Proc. IEEE Symp. on Research in Security and Privacy (S&P 1998)*, pages 66–92, Oakland, CA, USA, May 1998. IEEE.

Krishna Sampigethaya, Leping Huang, Mingyan Li, Radha Poovendran, Kanta Matsuura, and Kaoru Sezaki. CARAVAN: Providing location privacy for VANET. In *Proc. Embedded Security in Cars (ESCAR 2005)*, pages 29–37, Tallinn, Estonia, July 2005.

Andrei Serjantov and George Danezis. Towards an Information Theoretic Metric for Anonymity. In *Proc. 2nd Int. Symp. on Privacy Enhancing Technologies (PETS 2002)*, LNCS 2482, pages 41–53, San Francisco, CA, USA, April 2002. Springer. ISBN 978-3-540-00565-0, 978-3-540-36467-2.

Stefaan Seys and Bart Preneel. ARM: Anonymous Routing Protocol for Mobile Ad hoc Networks. *Int. Journal of Wireless and Mobile Computing*, 3(3):145–155, October 2009.

Claude Elwood Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423 & 623–656, October 1948.

Min Shao, Yi Yang, Sencun Zhu, and Guohong Cao. Towards Statistically Strong Source Anonymity for Sensor Networks. In *Proc. 27th Conf. on Computer Communications (INFOCOM 2008)*, pages 466–474, Phoenix, AZ, USA, April 2008. IEEE.

Elaine Shi, T-H. Hubert Chan, Eleanor G. Rieffel, Richard Chow, and Dawn Song. Privacy-Preserving Aggregation of Time-Series Data. In *Proc. 18th Annu. Network & Distributed System Security Symp. NDSS'2011*, volume 2, page 4, San Diego, CA, USA, February 2011.

Vitaly Shmatikov. Probabilistic Analysis of Anonymity. In *Proc. 15th IEEE Computer Security Foundations Workshop (CSFW-15)*, pages 119–128, Cape Breton, Canada, June 2002. IEEE.

R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J-P Hubaux. Quantifying Location Privacy. In *Proc. 2011 32nd IEEE Symp. on Security and Privacy (S&P 2011)*, pages 247–262, Oakland, CA, USA, May 2011. IEEE.

Reza Shokri, Julien Freudiger, and Jean-Pierre Hubaux. A Unified Framework for Location Privacy. In *Proc. 3rd Symp. on Hot Topics in Privacy Enhancing Technologies (HotPETs 2010)*, Berlin, Germany, July 2010a.

Reza Shokri, Carmela Troncoso, Claudia Diaz, Julien Freudiger, and Jean-Pierre Hubaux. Unraveling an Old Cloak: K-anonymity for Location Privacy. In *Proc. 9th ACM Workshop on Privacy in the Electronic Society (WPES 2010)*, pages 115–118, Chicago, Illinois, USA, October 2010b. ACM. ISBN 978-1-4503-0096-4.

Jordi Soria-Comas and Josep Domingo-Ferrert. Differential Privacy via t-Closeness in Data Publishing. In *Proc. 11th Annu. Conf. on Privacy, Security and Trust (PST2013)*, pages 27–35, Tarragona, Spain, July 2013. IEEE.

Sandra Steinbrecher and Stefan Köpsell. Modelling Unlinkability. In *Proc. 3rd Int. Workshop on Privacy Enhancing Technologies (PET 2003)*, LNCS 2760, pages 32–47, Dresden, Germany, March 2003. Springer. ISBN 978-3-540-20610-1, 978-3-540-40956-4.

Latanya Sweeney. k-Anonymity: A Model for Protecting Privacy. *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, October 2002. ISSN 0218-4885.

Paul Syverson. Why I'm Not an Entropist. In *Proc. 17th Int. Workshop on Security Protocols*, LNCS 7028, pages 213–230, Cambridge, UK, April 2013. Springer. ISBN 978-3-642-36212-5, 978-3-642-36213-2.

Kurt Thomas, Chris Grier, and David M. Nicol. unFriendly: Multi-party Privacy Risks in Social Networks. In *Proc. 10th Int. Symp. on Privacy Enhancing Technologies (PETS 2010)*, LNCS 6205, pages 236–252, Berlin, Germany, July 2010. Springer. ISBN 978-3-642-14526-1, 978-3-642-14527-8.

Gergely Tóth, Zoltán Hornák, and Ferenc Vajda. Measuring Anonymity Revisited. In *Proc. 9th Nordic Workshop on Secure IT Systems (Nordsec 2004)*, pages 85–90, Espoo, Finland, November 2004.

NN Vorob'ev. Infinite antagonistic games. In *Game Theory*, volume 7 of *Applications of Mathematics*, pages 56–89. Springer, 1977.

Nevena Vratonjic, Kévin Huguenin, Vincent Bindschaedler, and Jean-Pierre Hubaux. How Others Compromise Your Location Privacy: The Case of Shared Public IPs at Hotspots. In *Proc. 13th Int. Symp. on Privacy Enhancing Technologies (PETS 2013)*, LNCS 7981, pages 123–142, Bloomington, IN, USA, July 2013. Springer. ISBN 978-3-642-39076-0, 978-3-642-39077-7.

Isabel Wagner. Genomic Privacy Metrics: A Systematic Comparison. In *36th IEEE Symposium on Security and Privacy (S&P): 2nd International Workshop on Genome Privacy and Security (GenoPri'15)*, San Jose, CA, May 2015.

Isabel Wagner and David Eckhoff. Privacy Assessment in Vehicular Networks Using Simulation. In *Proc. Winter Simulation Conf. (WSC '14)*, Savannah, GA, USA, December 2014.

Ke Wang and Benjamin Fung. Anonymizing Sequential Releases. In *Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and data Mining (KDD'06)*, pages 414–423, Philadelphia, PA, USA, August 2006. ACM.

Ke Wang, Benjamin CM Fung, and S. Yu Philip. Handicapping Attacker's Confidence: An Alternative to k-Anonymization. *Knowledge and Information Systems*, 11(3):345–368, April 2007.

Rui Wang, XiaoFeng Wang, Zhou Li, Haixu Tang, Michael K. Reiter, and Zheng Dong. Privacy-Preserving Genomic Computation Through Program Specialization. In *Proc. 16th ACM Conf. on Computer and Communications Security (CCS'09)*, pages 338–347, Chicago, IL, USA, November 2009. ACM. ISBN 978-1-60558-894-0.

Alan Westin. *Privacy and Freedom*. Atheneum, 1967.

Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, and Ke Wang. (α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In *Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and data Mining (KDD'06)*, pages 754–759, Philadelphia, PA, USA, August 2006. ACM.

Matthew Wright, Micah Adler, Brian Neil Levine, and Clay Shields. An Analysis of the Degradation of Anonymous Protocols. In *Proc. Network and Distributed System Security Symp. (NDSS'02)*, volume 2, pages 39–50, San Diego, CA, USA, February 2002.

Matthew Wright, Micah Adler, Brian Neil Levine, and Clay Shields. Defending Anonymous Communications Against Passive Logging Attacks. In *Proc. IEEE Symp. on Research in Security and Privacy (S&P 2003)*, pages 28–41, Oakland, CA, USA, May 2003. IEEE.

Xiaokui Xiao and Yufei Tao. Personalized Privacy Preservation. In *Proc. 2006 ACM SIGMOD Int. Conf. Management of Data (SIGMOD 2004)*, pages 229–240, Chicago, IL, USA, June 2006. ACM.

Xiaokui Xiao and Yufei Tao. M-invariance: Towards Privacy Preserving Re-Publication of Dynamic Datasets. In *Proc. ACM SIGMOD Int. Conf. on Management of data (SIGMOD '07)*, pages 689–700, Beijing, China, June 2007. ACM. ISBN 978-1-59593-686-8.

Toby Xu and Ying Cai. Feeling-based Location Privacy Protection for Location-based Services. In *Proc. 16th ACM Conf. on Computer and Communications Security (CCS'09)*, pages 338–347, Chicago, IL, USA, November 2009. ACM. ISBN 978-1-60558-894-0.

Yi Yang, Min Shao, Sencun Zhu, Bhuvan Urgaonkar, and Guohong Cao. Towards Event Source Unobservability with Minimum Network Traffic in Sensor Networks. In *Proc. 1st ACM Conf. on Wireless Network Security (WiSec'08)*, pages 77–88, Alexandria, VA, USA, April 2008. ACM. ISBN 978-1-59593-814-5.

Yuhao Yang, Jonathan Lutes, Fengjun Li, Bo Luo, and Peng Liu. Stalking Online: On User Privacy in Social Networks. In *Proc. 2nd ACM Conf. on Data and Application Security and Privacy (CODASPY'12)*, pages 37–48, San Antonio, TX, USA, May 2012. ACM. ISBN 978-1-4503-1091-8.

Danfeng Yao, Keith B. Frikken, Mikhail J. Atallah, and Roberto Tamassia. Private Information: To Reveal or Not to Reveal. *ACM Transactions on Information and Systems Security*, 12(1): 6:1–6:27, October 2008. ISSN 1094-9224.

Fei Yu, Stephen E. Fienberg, Aleksandra B. Slavković, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141, August 2014. ISSN 1532-0464.

Sherali Zeadally, Al-Sakib Khan Pathan, Cristina Alcaraz, and Mohamad Badra. Towards Privacy Protection in Smart Grid. *Wireless Personal Communications*, 73(1):23–50, November 2013. ISSN 0929-6212, 1572-834X.

Kehuan Zhang, Xiaoyong Zhou, Yangyi Chen, XiaoFeng Wang, and Yaoping Ruan. Sedic: Privacy-aware Data Intensive Computing on Hybrid Clouds. In *Proc. 18th ACM Conf. on Computer and Communications Security (CCS'11)*, pages 515–526, Chicago, IL, USA, October 2011. ISBN 978-1-4503-0948-6.

Lei Zhang, Sushil Jajodia, and Alexander Brodsky. Information Disclosure Under Realistic Assumptions: Privacy Versus Optimality. In *Proc. 14th ACM Conf. on Computer and Communications Security (CCS'07)*, pages 573–583, Alexandria, VA, USA, October 2007a. ACM. ISBN 978-1-59593-703-2.

Qing Zhang, Nick Koudas, Divesh Srivastava, and Ting Yu. Aggregate Query Answering on Anonymized Tables. In *Proc. IEEE 23rd Int. Conf. on Data Engineering (ICDE 2007)*, pages 116–125, Istanbul, Turkey, April 2007b. IEEE.

Xiaoyong Zhou, Bo Peng, Yong Fuga Li, Yangyi Chen, Haixu Tang, and XiaoFeng Wang. To Release or Not to Release: Evaluating Information Leaks in Aggregate Human-Genome Data. In *Proc. 16th Symp. on Research in Computer Security (ESORICS 2011)*, LNCS 6879, pages 607–627, Leuven, Belgium, September 2011. Springer.